CIP Competitiveness and Innovation Framework Programme
CIP-Pilot Actions, 2007-2013
CIP-ICT-PSP-2012-6

Project **CIP-Pilot 325101 / OpenScienceLink**
Deliverable **D2.1**
Distribution **Public**

http://opensciencelink.eu

Requirements, Use Cases and KPIs

Authors: Adomas Bunevicius, Todor Tagarev, Petya Tagareva, Vassiliki Andronikou, Stathis Karanastasis, Costas Pantos, Iordanis Mourouzis, Giorgio Iervasi, Sara Hugelier, Matthias Zschunke, George Tsatsaronis

Status: Final (Version 1.0)

September 2013

Project

Project ref.no.	CIP-Pilot 325101
Project acronym	OpenScienceLink
Project full title	Open Semantically-enabled, Social-aware Access to Scientific Data
Project site	http://opensciencelink.eu
Project start	February 2013
Project duration	3 years
EC Project Officer	Kirsti Ala-Mutka

Deliverable

Deliverable type	Report
Distribution level	Public
Deliverable Number	D2.1
Deliverable title	Report on Stakeholders, Main Use Cases, KPIs and Data Sources Assessment
Contractual date of delivery	M5 (June 2013)
Actual date of delivery	September 2013
Relevant Task(s)	WP2/Tasks 2.1, 2.2, 2.3 and 2.4
Partner Responsible	LUHS
Other contributors	NTUA, LUHS, NKUA, KU Leuven, CNR, Procon, TI
Number of pages	139
Author(s)	Adomas Bunevicius, Todor Tagarev, Petya Tagareva, Vassiliki Andronikou, Stathis Karanastasis, Costas Pantos, Iordanis Mourouzis, Giorgio Iervasi, Sara Hugelier, Matthias Zschunke, George Tsatsaronis
Status & version	Final
Keywords	D2.1, User Requirements, Use Cases, KPIs

Executive Summary

The current deliverable reports in detail the requirements engineering processes associated with all stakeholders of the OpenScienceLink platform. In addition, this deliverable illustrates requirements associated with the interfacing to multiple openly accessible scientific repositories (including the repositories to be used during the pilot operations and more). Finally, the deliverable presents the main use cases associated with the OpenScienceLink pilot services, along with detailed Key Performance Indicators (KPIs) for their monitoring and auditing.

For the purposes of the aforementioned analysis, the document is structured into 6 sections. Section 1 summarizes the OpenScienceLinik project landscape, and discusses the objectives of the project which set the basis for the OpenScienceLink platform requirements. Section 2 lists and analyses the requirements for the OpenScienceLink platform from the view of all involved stakeholders, i.e., researchers, evaluators, publishers, funders, and journalists/press. In addition, special focus is given to the requirements stemming from the legal analysis of the platform services. Section 3 illustrates the use cases across all five OpenScienceLink platforms, which set the functional requirements for the platform implementation and constitute the basis for the technical requirements. Section 4 focuses on the requirements associated with the underlying/used content sources from the OpenScienceLink platform. Section 5 presents the Key Performance Indicators associated with the monitoring and auditing of the OpenScienceLink project progress and evaluation of the resulting platform. Finally, Section 6 summarizes the contents of the deliverables and concludes.

Contents

List of Figures

List of Tables

1 Introduction

As detailed in the project's Description of Work (OpenScienceLink Consortium, 2013), the core objective of the OpenScienceLink project is to introduce and pilot a holistic approach to the publication, sharing, linking, review and evaluation of research results, based on the open access to scientific information. Motivated by the EC recommendations and respective policies towards supporting experiments with open access to scientific information (European Commission, 2013), including experiments exploring new paradigms for rendering, querying, mining, linking and evaluating scientific content, OpenScienceLink aims at becoming one such vehicle that will allow researchers globally to publish their scientific results, findings and thoughts in an open access manner. Since scientific information refers to the results of scientists' or scholars' research work (in particular scientific articles and associated datasets, monographs) in the EU Member States or associated countries, within OpenScienceLink the consortium aims at developing data journals and linking data to scientific articles, finding new ways to peer review articles and the associated data, opening up new possibilities for data mining of journal articles and of research data and developing new research evaluation systems taking into account the way in which results are made available to other researchers (e.g., complementing the ISI/Impact Factor method). The expected impact is the opening to wider access of scientific information and the contribution towards creating new ways to review and evaluate scientific information based on publicly accessibly data and statistics. Overall, the vision under which the OpenScienceLink project operates, is the one illustrated in Figure 1 which describes the general principles and goals of the EC Horizon 2020 programs that aim at transforming science through ICT tools, networks and media, to make research more open, global, collaborative, creative and closer to society. The vision is that of 'democratizing' research, and transforming it into transparent and replicable.

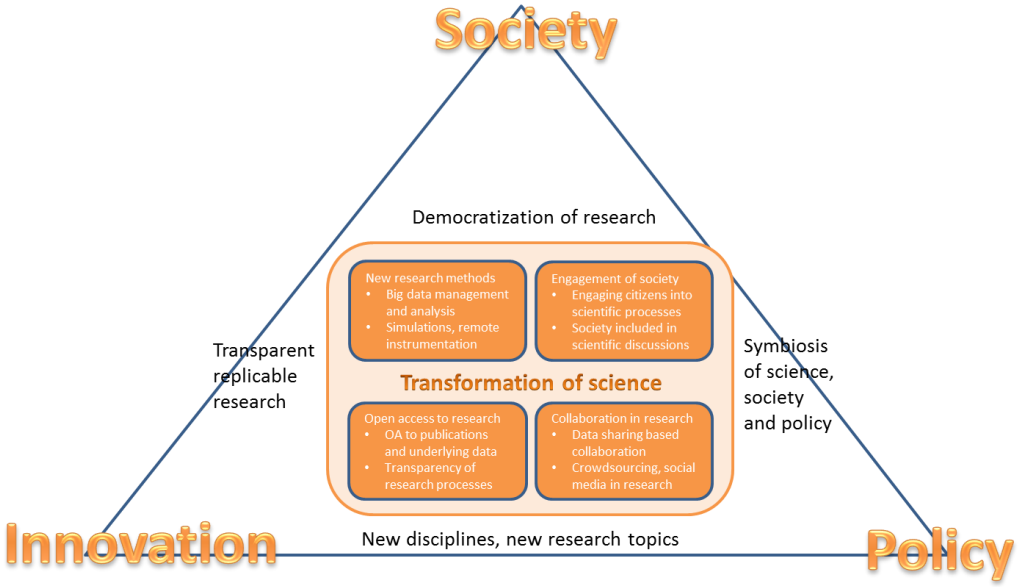


Figure 1: The vision of 'democratizing' research.

Towards this end, OpenScienceLink will pilot a range of novel services that could alleviate the lack of structured data journals and associated data models, the weaknesses of the review process, the poor linking of scientific information, as well as the limitations of current research evaluation metrics and indicators. Five pilot services will be integrated and piloted in particular:

- Data journals development based on semantically-enabled research dynamics detection,
- A novel open, semantically-assisted peer review process,
- A services for detection and analysis of research trends,
- Services for Dynamic researchers' collaboration based on non-declared, semantically-inferred relationships, and
- A set of scientific field-aware, productivity- and impact-oriented enhanced research evaluation services.

These services will be developed over the OpenScienceLink platform, which will be based on the semantic and social networking capabilities of background FP7 projects, as well as of the popular GoPubMed search engine[1]. The OpenScienceLink services will be piloted with the active participation of over 1200 researchers from the consortium organizations. OpenScienceLink has already established a group of external users/stakeholders that will contribute additional users/researchers in the scope of the validation process, while also engaging in the sustainable use of the services.

OpenScienceLink will also study the business potential of open access paradigms, through investigating and pursuing multiple business models including author fees, hard copy sales, advertisements, sponsorship, as well as subscription based models. Furthermore, as part of its holistic approach, OpenScienceLink will devise and validate a legal framework for regulating and reusing open scientific data.

The focus of WP2 is to elicit, collect, analyze and document requirements associated with the OpenScienceLink platform and services, including requirements about the overall OpenScienceLink paradigm to open access services. The main objectives of this work package are:

- To collect and analyze requirements from all stakeholders associated with the development, integration, deployment and operation of the OpenScienceLink platform and pilot services.
- To collect and analyze requirements associated with the collection and use of multiple openly accessible repositories, on the basis of appropriate access interfaces.
- To specify the main use cases associated with the OpenScienceLink platform and pilot services.
- To specify a number of Key Performance Indicators (KPIs) associated with the planning and implementation of the OpenScienceLink platform and services, as well as with their sustainability and wider use.

The current deliverable constitutes the main document that details the requirements engineering processes associated with all stakeholders of the OpenScienceLink platform. In addition, this deliverable describes the requirements associated with the interfacing to multiple openly accessible scientific repositories (including the repositories to be used during the pilot operations and more). Finally, it details the main use cases associated with the OpenScienceLink pilot services, along with detailed KPIs for their monitoring and auditing.

[1] http://www.gopubmed.org/web/gopubmed/

2 OpenScienceLink Requirements

The aim of work package 2 within the OpenScienceLink project is to deliver a set of clearly documented and coded requirements of all potential stakeholders in the holistic approach to the publication, sharing, linking, review and evaluation of research results, based on the open access to scientific information. While the research team strived to define a comprehensive list of requirements, the focus was on stakeholders requirements associated with the development, integration, deployment and operation of the OpenScienceLink platform and pilot services, that will be demonstrated in the project life cycle.

The OpenScienceLink The identifies five main groups of stakeholders:

1. "Researchers": Scientists, Researchers, Scholars;
2. "Evaluators": Evaluators, Referees, Reviewers;
3. Publishers;
4. "Funding Agencies": Corporate Sponsors, Venture Capitals/Funds, Government Funding Agencies (possibly adding inter-/ supra-governmental agencies, such as the EU);
5. Journalists, Press and, thus, wider society.

To elicit requirements, the OpenScienceLink partners used a number of modalities:

1. Getting into direct contact with stakeholders – within and outside the consortium, e.g., through the partners business networks;
2. Consolidating their own experience;
3. Analysing the experience of past projects and related ongoing projects where information is available;
4. Analysing official documents, including EU documents;
5. Reviewing the related professional literature.

Where possible, partners tried to identify trends and distinguish current/near-term and future requirements. This chapter of the report starts with a review of the requirements capturing process. Then, it presents the findings on requirements of the five groups of stakeholders and common 'non-functional' requirements. Next, the chapter provides review of the requirements from a legal point of view. The final section of the chapter structures these requirements per OpenScienceLink pilot.

2.1 The Requirements Capturing Process

2.1.1 Requirements types and definition process

One of the objectives of OpenScienceLink Work Package 2 is to collect and analyse requirements from all stakeholders associated with the development, integration, deployment and operation of the OpenScienceLink platform and pilot services. A project of the scope and complexity of OpenScienceLink has a variety of stakeholders with their specific requirements.

Requirements are of two main types (MITRE, 2012a):

- *Functional*, associated with the capability or application needed to directly support the users' accomplishment of their mission and tasks, and
- *Non-functional*, which are typically implicit and technical in nature and emerge as system requirements to satisfy the users' functional needs, e.g., quality of service, availability, timeliness, and accuracy.

Requirements to complex systems are elicited based on users' informal narratives, observation of the user environment, or capturing user responses to targeted questions (MITRE, 2012a). Further, MITRE system engineers distinguish three process models in eliciting requirements:

- Waterfall;
- Spiral;
- Agile.

Fellows of the International Council on Systems Engineering (INCOSE) have reached a consensus on the process of satisfying stakeholders' needs and requirements in a high quality, trustworthy, cost efficient and schedule compliant manner throughout a system's entire life cycle (INCOSE, 2006). This process, known as SIMILAR, is comprised of the seven tasks—**S**tate the problem, **I**nvestigate alternatives, Model the system, **I**ntegrate, **L**aunch the system, **A**ssess performance, and **R**e-evaluate—performed in a parallel and iterative manner (Bahill and Gissing, 1998).

One should be able to trace mandatory and preference requirements to the problem statement, elaborated in the first phase of the process and describing the top-level functions that the OpenScienceLink platform must perform. Often, this problem statement is in the form of a *mission statement*, a *concept of operations* or a *description of deficiencies* that should be overcome. The problem statement expresses stakeholders' requirements in functional or behavioural terms (INCOSE, 2006).

Requirements are elicited and refined within a process that is parallel and iterative, and not sequential. In the OpenScienceLink project this is reflected by re-opening Work Package 2 to refine requirements after the first two phases of deployment, validation and evaluation of OpenScienceLink pilot services.

2.1.2 Characteristics of a good requirements statement

A good statement of requirements can typically be characterised as follows (MITRE, 2012b):

First, a requirement must be *traceable* to an operational need and attributable to an authoritative source, e.g. a document or a person. Once defined, it receives a unique identifier allowing the software design, code, and test procedures to be precisely traced back to the requirement.

Requirement definition should be *unambiguous*. A good practice is to test the wording of the requirement from the perspectives of different stakeholders and check whether it can be interpreted in multiple ways. Vague, general statements must be avoided. This will allow testing the requirements and demonstrating that they are satisfied by the end product or service. Descriptions need to be *clear*, *specific* and *singular*.

Definitions need to be *measurable*, either quantitatively or qualitatively. Typical categories of measures are:

- Measures of Effectiveness (MOEs), i.e. measures of mission success from stakeholders point of view;
- Measures of Performance (MOPs), used to determine whether the system meets performance requirements necessary to satisfy the MOE.
- Key Performance Parameters (KPPs) or Indicators (KPIs) defined by stakeholders as measures of minimal and critical system or project performance and level of acceptance.

Requirements must be *uniquely identified*, *consistent* and *compatible* with each other.

Requirements need to be *feasible*, i.e. they are attainable in terms of technology, cost, and schedule. If a requirement cannot be implemented, it should be removed from the statement.

Finally, the specification of requirements should be *design-free*, reflects *what* the system need to accomplish.

2.1.3 Good practices in requirements definition

Ivy Hooks (2005) presents a study on deficiencies in specifying requirements. The following are the top among identified defects:

- Incorrect information;
- Omissions;
- Ambiguities;
- Poorly written;
- Misplaced;
- Implementation or operations-related, and not design-free definitions.

Further, Hooks provides an advice on how to prevent requirements defects (see Table 1).

Requirement specification defects	**Ways of prevention**
Incorrect Information	Complete scope Operational concepts Rationale Include stakeholders
Omissions	Identify necessary reiteration Use a standard outline/checklist
Ambiguities	Use standards Validate
Poorly Written	Use simple format Use editor
Misplaced	Standard outline (template)
Implementation or Operations	Ask “Why?” Ask “What do you want to verify?

Table 1: Common defects in specifying requirements and ways of prevention.

Below is an excerpt of best practices in requirements definition identified by MITRE systems engineers (MITRE, 2012b):

- Baseline and agree. Developing requirements is usually a collaborative activity, involving users, developers, maintainers, integrators, etc., so avoid placing the responsibility of requirements analysis solely on one stakeholder. When all team members acknowledge a set of requirements is done, this is called a baseline (recognising that definitions will evolve).
- Requirements analysis is an iterative process. At each step, the results must be compared for traceability and consistency with stakeholders' requirements, and then verified with users, or go back into the process for further analysis, before being used to drive architecture and design.
- Special attention is to be paid to interface requirements. Requirements must clearly capture all the interactions with external systems and the external environment so that boundaries are clear.
- Be flexible. To balance out rigidness of baselining requirements, a development team should consider what constitutes a "change of requirements" as distinguished from a valid interpretation of requirements. The key is to find a balance between adherence to a baseline and sufficient flexibility.
- Use templates and tools that suit your needs.

The next section provides the first definition of the set of stakeholders' requirements to the OpenScienceLink platform and associated activities.

2.2 Stakeholders' Requirements

2.2.1 Researchers Requirements

Open access to scientific information is by all means at the forefront of science development and worldwide initiatives of large organizations and bodies, such as the EC and UNESCO, to name a few. Open access is expected to provide great benefits to researchers, innovators, teachers, students, media, professionals, and the general public. According to a recent study, the majority of the researchers believe that access to publicly available services and tools will lead to a significant increase in scientific networking, collaboration activities, and increased research competitiveness in the next decade (Lery and Bressler, 2007). In the same study it was reported, following a large scale analysis of researchers' requirements on network-related services and tools, that the majority of the researchers in almost all science fields access frequently digital libraries and remote databases, and the prediction is that among the researchers who do not use yet such services tools, approximately 40% will do so in the next coming years. Most interestingly, digital libraries and services were found to be accessed for a wide variety of processes, which include the following areas:

- Literature searches of journals, magazines and e-books
- Accessing conference proceedings
- Accessing software toolboxes
- Receiving e-mail alerts from journals
- Accessing databases and digital archives of data

Among others, the majority of researchers find that access to the aforementioned services affects 'very positively' or 'positively' the access to the information needed for own research, the ability to obtain information on meetings and conferences, and the ability to obtain access to new fields of research; all of them promoting research advances, collaborations, and competitiveness. In addition, the majority of researches find that over the next decade a number of research areas will develop further in high pace, including the study domain of the OpenScienceLink project, e.g., the biomedical domain, in the extent to which collaboration would

take place with other researchers abroad. In addition, more than 80% of the respondents in the same study expected increases in large collaborative project work and expanded collaborations with domestic researchers.

From the aforementioned previously reported results, and taking into account the opinions and views of the researchers in the OpenScienceLink consortium, the OpenScienceLink pilots should have as axes the following requirements:

F_Rsr_1 The researchers should be able to register and create a profile based on their academic/research interests.

F_Rsr_2 The profile should be used to automatically suggest possible research collaborations at the international and domestic level.

F_Rsr_3 The registered researchers should be able to see the advances (emerging fields and trendiness) in their fields of interest.

F_Rsr_4 The registered researchers should be able to see the top authors, research institutes and countries in the field of their study and interest based on the evaluation metrics of their preference.

F_Rsr_5 The registered researchers should be able to share and discuss scientific news.

F_Rsr_6 The registered researchers should be able to comment/discuss previously published research works or datasets.

F_Rsr_7 The registered researchers should be able to upload their scientific data for sharing and publication under an open-access data journal.

F_Rsr_8 All researchers should be able to search and have access to as much as possible scientific information, such as research papers, data sets, and data repositories.

F_Rsr_9 The search results should be presented in an organized manner, following the classification of research works under pre-specified topics.

F_Rsr_10 Search capabilities should be enhanced with the ability to filter results based on specific keywords that represent main concepts in the domain of interest (semantic search).

F_Rsr_11 Researchers should be given the ability to peer review research works and openly discuss and express opinions on the reviews and the review results.

F_Rsr_12 Researchers should be able to have access to the results of research evaluation at all levels – journal, individual researcher, research group, research institute/university, country.

F_Rsr_12.1 Researchers could be able to choose the metrics to be applied for evaluating journals, individual researchers, research groups, research institute/universities, countries.

F_Rsr_12.2 Researchers could be able to rank the top journals, individual researchers, research groups, research institute/universities, countries in a specific domain based on the evaluation metrics of their preference.

F_Rsr_13 The evaluation criteria and methodology should be publicly available and should be ideally based on measurements that may be drawn by accessing publicly accessible data sources, or publicly accessible statistical information and reports.

F_Rsr_14 The registered researchers should be able to receive notifications based on pre-selected criteria with regards to newly published research work or data sets in their fields of interest.

F_Rsr_15 Researchers are presented with automatically-generated proposed collaborations with other researchers as well as research groups and communities.

F_Rsr_16 The suggested collaborations should be generated based on the degree of relevance of the research topics and fields they work on.

F_Rsr_17 Proposed collaborations should take into consideration their former and existing collaborations and social-media declared relationships.

F_Rsr_18 The suggested collaborations are presented with a justification for the researcher to be able to evaluate them:

F_Rsr_18.1 A list of the most relevant publications of the researcher who is suggested for collaboration

F_Rsr_18.2 The scientific topics, fields and/or areas in which they share common interests with the researcher

F_Rsr_18.3 Potentially a percentage indicating how strongly suggested the collaboration is

F_Rsr_19 The suggested collaborations are communicated to the researchers either through the platform or via e-mail.

F_Rsr_20 Rising research topics as well as automatically formulated groups of researchers who share common interests could form the basis for automatically proposed research communities to which the researchers are invited to participate and/or lead based on their excellence in the field.

F_Rsr_21 Researchers could be able to receive automatic updating about their own evaluation, including their research work and group they belong to.

F_Rsr_22 Researchers could be able to follow the evaluation of specific researchers, journals, papers, research groups, academic institutions and countries.

2.2.2 Evaluators Requirements

Spreading novel ideas is a major goal of the scientific community. Evaluation processes are in place to ensure the quality of publications and assure the compliance with scientific standards. For a quality journal it is necessary to judge whether a potential publication covers a novel and substantial subject. The vast majority of the scientific journals facilitate a peer review process. Peer reviewers or referees are independent scientific individuals capable to judge a certain work in a specific field.

Funding is another topic where evaluation has to take place. Funding agencies need to decide whether a specific funding proposal is worth spending the money. They have to dedicate and target budgets to upcoming research areas. They need to ensure that proposals exploit the results from other work already done or funded.

Evaluators of scientific work, as for example reviewers, need to get on track in terms of the topic of a given publication to be evaluated. They need to get in touch and meet with other scientists to discuss the different aspects of the prospective work. They need to know the potential ways the corresponding scientific community is going to take. And they need at least to overview the margin fields of their scientific area. As the communication potentials of the internet are facilitating a rapid growth in scientific literature and brings together ever-growing communities of scientists, evaluators face a tough work to identify relevant resources for their judgement.

There is a necessity to support the evaluators with focused material to make their work effective and efficient. It is important that evaluators get a fast access to the information they need to judge the quality of a scientific article or a transmitted dataset. From the OpenScienceLink platform the evaluators, such as reviewers, expect the following capabilities:

F_Evl_1 For an uploaded dataset, retrieve the metadata and the data.

F_Evl_2 For a publication, upload the publication for further processing.

F_Evl_3 Retrieve related bibliography covering the topics or the datasets

F_Evl_4 Retrieve cited bibliography.

F_Evl_5 Retrieve information on separate and collaborating authors in the same field / on the same topics.

F_Evl_6 Retrieve related datasets to suggest comparisons.

F_Evl_7 Accept or reject an invitation for review.

F_Evl_8 View reviews of other evalutators on already published work.

F_Evl_9 Maintain the reviews within the platform for further processing, e.g., discussion with colleague researchers.

F_Evl_10 Prepare the review in a provided form and submit the review to the authors / researchers, editor and funder.

F_Evl_11 Retrieve information on the scientific direction and trends of a specific topic.

F_Evl_12 Evaluators are anonymous w.r.t. the authors / researchers and colleague evaluators.

F_Evl_14 Evaluators maintain a profile within the platform to enable recruitment by editors and funders.

F_Evl_14.1 The profile contains concrete working topics.

F_Evl_14.2 The profile contains the publication list of the evaluator with direct links to the publications.

F_Evl_15 Evaluators provide a valid email address, so they can be contacted by editors and funders.

F_Evl_16 Evaluators employed by funding agencies have access to already funded work in the respective fields.

F_Evl_17 Evaluators retrieve other publications of the author.

2.2.3 Publishers Requirements

In the age on internet and social networks, publishers continue to play a key role in providing access to research results with quality assurance, based on rigorous assessment.

This is increasingly the case in the field of open access journals. The Directory of Open Access Journals (DOAJ) already features 9817 journals published in 120 countries and over 1.13 million articles.[2] David Lewis predicts that 'Gold Open Access'—a model of publishing where all articles of a journal are freely available at the time of publication—"could account for 50 percent of the scholarly journal articles sometime between 2017 and 2021, and 90 percent of articles as soon as 2020 and more conservatively by 2025" (Lewis, 2012: 493).

This section of the report presents publishers' requirements to the "holistic approach to the publication, sharing, linking, review and evaluation of research results, based on the open access to scientific information" (DOW, 2013: 3).

These requirements were elicited by review, analysis and consolidation of the experience of the publisher in the consortium, review the related professional literature and guidance provided by the Directory of Open Access Journals (DOAJ), the Publishers International Linking Association

[2] See www.doaj.org. Data as of 5 July 2013.

(PILA), etc., and analysis of the approach of other publishers as declared in their related policies and statements.[3]

From the perspective of the publisher in the consortium—an SME that combines publishing with research and consultancy services—the publisher's goals within the "holistic approach to the publication, sharing, linking, review and evaluation of research results, based on the open access to scientific information" can be defined as follows:

> *Goal 1*: Facilitate wide and open access to scientific information by publishing one or more open-access journals presenting research data that will be sustainable in mid- to long-term.
>
> *Goal 2*: Strengthen the competitive advantages of the publishing company.

Rationale for Goal 1: Open access to scientific and scholarly journals promotes their increased usage and impact. Harnad and Brody (2004) discovered "dramatic citation advantages for open access" journals. Furthermore, the Berlin Declaration (2003) aimed to disseminate knowledge via Internet and make it widely and readily available to society. 435 universities, academies, and library associations are already signatories to the Berlin Declaration (Max Plank Society, 2013). Some signatories enhance their open access policy by requesting that faculty members—authors or co-authors of future scholarly articles—make these articles available for free through the institutional digital repository (Oregon State University, 2013).

With the trend of rapidly expanding open access to scientific publications, newcomers to the field of scientific publishing—both publishers and new journals—would likely find it impossible to implement traditional models of paid subscriptions. They have instead to find ways to sustain the publication of a journal that is freely available to its readers. Within the OpenScienceLink concept, the publication of one or more open access data journals needs to provide revenues that, in a multi-year frame, exceed the expenses for creating and maintaining the journals.

Other publishers, in particular large ones, may be using more traditional statements in defining there goals, involving decision criteria (and then performance measures) in terms of profitability. One example is to consider potential ROI (Return on Investment),[4] where

Return on Investment = (Gain from investment – Cost of investment) / Cost of investment

We were not able to identify sources addressing the reliability of using measures of profitability for planning purposes in the field of open access publishing. One reason is that business models are rapidly evolving, without one clearly prevailing over others at this point.

Thus, the first goal is defined in sustainability terms, where sustainability is sought in a multi-year framework. Hence, in mid- to long-term a publisher would expect that revenues exceed the expenditures in running one or more open access data journals.

Before addressing publisher requirements *per se*, we will briefly touch on some potential business models and the anticipated understanding of their feasibility. The actual study and elaboration of business models is subject of research in tasks T 8.4 and T 9.3 of the OpenScienceLink project. An initial list of key features of such models includes:

a. Paid access to new issues (open access is provided after a certain delay, e.g. six months or until a new issue is published; that means free access to old issues);[5]

---

[3] Without getting into direct contact with them /in an attempt to preserve the anticipated competitive advantage/.

[4] Among the related metrics are Rate of Return (RoR, also known as 'rate of profit' or sometimes just 'return,' is the ratio of money gained or lost (whether realized or unrealized) on an investment relative to the amount of money invested), Return on Assets (RoA), Return on Net Assets (RoNA), Return on Capital (RoC), Return on Capital Employed (ROCE), and Return on Invested Capital (RoIC). See, for example, http://en.wikipedia.org/wiki/Return_on_investment.

[5] Some authors call this 'embargo period' or 'delayed Open Access.' However, this and the following model clearly contradict the journal selection criteria of DOAJ (Directory of Open Access Journals). See *DOAJ announces new selection criteria*, 12 June 2013, Retrieved from www.doaj.org/doaj?func=news&nId=303.

b. Free access to the html version (possibly excluding high resolution images and figures); paid access to the high resolution printable version, e.g. in high resolution pdf format;
c. Paid access to related services (to be specified in OpenScienceLink tasks T 8.4 and T 9.3);
d. Authors' fees: Charging the author for processing a submission and/or publishing a paper;
e. Institutional membership fees (in combination with the two modalities above);
f. Support by funding agencies;
g. Support by academic libraries, foundations, corporations, etc.;
h. Advertising (e.g. context sensitive placement of ads on the journal/article web page).

A feasible business model for publishing open access data journals may be based on a combination of two or more of these features. In addition, Open Access to research results is considered a disruptive innovation (Lewis, 2012) and different stakeholders react differently to this innovation.[6] For these reasons, it is likely that during the work of OpenScienceLink we will witness further proliferation and adaptation of business models and may not be in a position to reach a good understanding of what would be a best model (or best models), especially in longer term. Therefore, the OSL platform shall provide flexibility and allow efficient implementation of key features of a number of business models, e.g. article processing charges, hard copy sales, advertisements and sponsorships, institutional memberships, collaborative publishing, etc.

In defining requirements in terms of sustainability, a publisher needs to account for potential revenues and expenditures, related to the publication of an open access data journal. Hence, revenues and expenditures considerations form the basis for defining two major groups of requirements. Another group addresses the needs of key decisions, e.g. to launch a new data journal or not.

Goal 1-related requirements – Milestone decisions

The OpenScienceLink platform is expected to facilitate the analysis of publication needs and opportunities, accounting for the model of publication. Such analysis may result in milestone decisions, such as:

- Launch a new data journal;
- Refocus an existing data journal;
- Terminate the publication of an existing data journal.

To support effectively such decisions, the OpenScienceLink platform and ecosystem should have the capacity to:

F_Pbl_1 Identify trends, emerging research fields, and fields in decline;

F_Pbl_2 Maintain awareness of competitive publications (both traditional and open access journals and conference proceedings);

F_Pbl_3 Maintain awareness of potential collaborators, in particular organizers of relevant conferences;

F_Pbl_4 Recognize the relevance of own publications.[7]

By meeting these requirements, the OpenScienceLink platform will provide for faster and more reliable decisions on areas which could be covered by new journals based on the dynamics of

[6] For the reaction of the European Commission see "Open Access in FP7," Policy Initiatives, last update: 14 November 2012, http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1300.

[7] This is not currently relevant for the publisher in the OpenScienceLink, but might be necessary for potentially replicating the OpenScienceLink concept by a large publisher.

research in these fields, refocusing existing journals to match new research trends, and/or terminating the publication of an existing journal.

Another milestone decision is whether to publish the journal in the traditional paper format. Due to the nature of the journal, it needs to provide efficient access to the data, described in a particular publication, and related information sources. In this regard, the online publication has indisputable advantages over the paper format. Furthermore, the publication of books (to be sold) with in-depth review of results and trends may be seen as an added value service and, hence, part of the business model.

Goal 1-related requirements – 'Revenues'/ Effectiveness/

In the face of the inherent uncertainty of future open access publishing, a publisher can hardly define specific targets in terms revenue. It can be assumed, however, that prospective revenues will be strongly correlated to the prestige the journal will have in the professional community and other stakeholders.

The data journal /journals/ shall be attractive to contributors to the publication process, including:

- Authors /research groups;
- Editors; and
- Reviewers

No less important is that the data journal is attractive to its wider community of users, including:

- Researchers
- Funding agencies and corporate sponsors;
- Journalists and mass media.

In order to attract the listed stakeholders, a data journal should deal with state-of-the-art and perspective research fields and methods. This is treated by the first group of requirements above.

Second, it should include contributions of assured scientific quality, providing added value with no excessive delays. Key for providing relevant and timely publications of high quality is having in place an efficient, reliable and fast research evaluation process.[8]

The OpenScienceLink platform and ecosystem will facilitate this process primarily by cross-linking research fields and candidate editors, reviewers, and contributors.

In that respect, publishers' requirements to the platform can be defined as follows:

F_Pbl_5 Identify and prioritize potential members of the Editorial Board vis-à-vis initial and emerging research fields to be covered by the journal;

F_Pbl_6 Identify potential contributors to the journal;

F_Pbl_7 Identify and prioritize candidate reviewers per submitted contribution while eliminating potential conflict of interests;

F_Pbl_8 Facilitate the review process by:

F_Pbl_8.1 Identifying and providing access to milestone publications and most up-to-date publications in the respective research field;

F_Pbl_8.2 Facilitating a collaborative review process on the basis of a crowd sourcing paradigm;

[8] The technical aspects of the publication process are treated in the third group of Goal 1-related requirements below.

F_Pbl_8.3 Facilitating the review process through the provision of review forms which capture various evaluation aspects;

F_Pbl_9 Track the performance of reviewers in terms of:

F_Pbl_9.1 providing timely, complete, and professional reviews;

F_Pbl_9.2 avoiding potential conflicts of interest.

Third, the data journal is expected to be of high impact. In addition to the relevance, quality, and timeliness of publication, considered above, the impact depends on the ability of the publisher to reach out to potential contributors and the wider user community by utilising the potential of the OpenScienceLink platform and other means. The effective reach-out is based on high visibility of the data journal. Once the interest among stakeholder is triggered, it needs to be met by guaranteed long-term access to all journal publications and to be sustained.

The respective publishers' requirements are defined in the following manner:

F_Pbl_10 Provide high visibility of the data journal by

F_Pbl_10.1 Effective organisation of journal and article metadata, in particular the selection of keywords, development/adaptation and use of taxonomies, etc.

F_Pbl_10.2 Increasing journal page/site ranking (e.g. via advertising, search engine optimization, active 'push' to search engines, internet marketing services,[9] etc.);

F_Pbl_10.3 Making the journal and individual publications easy to find, e.g., by registering the journal with an ISSN agency, listing it and providing articles' metadata to directories, aggregators, indexing services, metadata harvesters, reference linking (PILA, 2012), 'cited by' linking (PILA, 2012), etc., as well as by the use of effective taxonomies /see requirement (a) above/ and website search tool;

F_Pbl_10.4 Present the journal in relevant social networks and regularly update the journal information;

F_Pbl_10.5 Present the journal in professional networks and provide regular updates (e.g., via listservs, sites, blogs, other specialized publications, exhibitions, etc.).

F_Pbl_11 Provide permanent access, quickly and in long-term:

F_Pbl_11.1 Establish permanent links to the journal, individual volumes, issues and articles, e.g. by using DOIs, handles, etc.;

F_Pbl_11.2 Provide immediate access to articles ready for publication ('pre-prints', i.e. before a whole journal issue is ready for publication);

F_Pbl_11.3 Provide online access to old issues and archives, e.g. via specialised archiving sites, in particular upon a decision to terminate the publication of a data journal;

F_Pbl_12 Sustain the interest in the journal by creating and nurturing a community of stakeholders (including readers/users of journal publications). Towards that purpose:

F_Pbl_12.1 Provide regular news/announcements service;

F_Pbl_12.2 Provide an opportunity to interested individuals to register and subscribe for getting news/announcements related to the journal; allow them to select particular fields of interest;

F_Pbl_12.3 Provide information to authors and other interested users, tailored to their particular interests;

[9] See, for example, www.bruceclay.com/eu.

F_Pbl_12.4 Provide opportunities to registered users to comment/ participate in discussions on journal publications, e.g. method, results, interpretation of results, etc.;

F_Pbl_12.5 Regularly inform the stakeholders on journal metrics and other information of interest (e.g. indications of recognition of journal publications and contributors);

F_Pbl_12.6 Track citations of published articles and inform respective authors.

To achieve and sustain its impact, the journal, as well as each individual contributor to the publication process, should maintain a high level of integrity.

F_Pbl_13 Enforce and maintain high level of integrity:

F_Pbl_13.1 Identify and eliminate cases of plagiarism, including auto-plagiarism;

F_Pbl_13.2 Seek guarantees that all authors, and only the authors, are acknowledged;

F_Pbl_13.3 Disclose (potential) conflicts of interest of one or more authors [10];

F_Pbl_13.4 Disclose (potential) conflicts of interest of reviewers;

F_Pbl_13.5 Manage access rights of journal website users (management; editors, reviewers, authors, 'interested readers,'[11] etc.); secure that all users of the journal website exercise their access rights according to the assigned privileges;

F_Pbl_13.6 Moderate discussions (prevent advertising in discussions, monitor use of language, decide on improper use of discussions and, when necessary, terminate privileges and ban users);

F_Pbl_13.7 Provide general security and integrity of the journal information and the supporting platform.

Goal 1-related requirements – Expenditures /Efficiency/

To increase the efficiency of journal management, i.e. to deliver a high quality product in a timely manner, the OpenScienceLink platform and ecosystem are expected to facilitate journal management by reducing costs in several ways.

A number of requirements relate to the efforts necessary to produce and maintain the journal.[12] In this respect, the OpenScienceLink platform and ecosystem are required to:

F_Pbl_14 Produce and disseminate calls for papers and other announcements (via listservs, dedicated sites, professional networks, etc.);

F_Pbl_15 Automate the publication management process, including:

F_Pbl_15.1 Submissions and communications with author/s;

[10] David Solomon (2008) provides the following definition of conflict of interest: "A conflict of interest exists when an author's financial interests or other opportunities for tangible personal benefit may compromise, or reasonably appear to compromise, the independence of judgment in the research or scholarship presented in the manuscript submission."

[11] For example, for commenting on a published paper and participating in discussions.

[12] Among the potential measures, to be explored in other work packages, is the required workload to manage a data journal, e.g. up to four PMs of a person of medium qualification (computer literate, undergraduate degree, with moderate English language abilities) per a quarterly journal per year. An alternative measure would be the effort (in PM) per published article or per submission.

F_Pbl_15.2 Review process (maintain a data base of reviewers, crowd sourcing, feedback, etc.);

F_Pbl_15.3 Tracking of revisions;

F_Pbl_15.4 Copyediting;

F_Pbl_15.5 Page setting and formatting;

F_Pbl_15.6 Exporting published articles in a number of formats, including formats for mobile reading devices;

F_Pbl_15.7 Uploading metadata and files to the journal website;

F_Pbl_15.8 Informing interested users on the publication of a new issue or an individual paper;

F_Pbl_15.9 Reducing the efforts/time necessary to include a journal issue in aggregators/indexing services, to assign a DOI, provide for 'cited by' linking and related functionalities;

F_Pbl_16 Automate the maintenance process, including:

F_Pbl_16.1 Efficient monitoring of discussions, support decision making on terminating privileges/banning users;

F_Pbl_16.2 Efficient tracking of usage, e.g. site/page visits and location of the user, referrals, downloads, etc.;

F_Pbl_17 Facilitate supporting activities, including:

F_Pbl_17.1 Efficient training of the persons responsible for publication management [13];

F_Pbl_17.2 Payment management (depending on the business model – by authors, sponsors/ funding agencies, 'pay per click', etc.);

F_Pbl_17.3 Efficient marketing.

A second group of requirements in this category address costs of external IT support, hosting, and back-ups and recovery services, i.e.:

F_Pbl_18 Low charges for hosting at (or use of) the OpenScienceLink platform [14];

F_Pbl_19 Provide robust storage and recovery of journal information, i.e., prevent loss and/or provide for efficient recovery of journal issues and other related data/ information.

The third group of requirements in this category relate to legal issues, including:

F_Pbl_20 Efficient handling of complaints;

F_Pbl_21 Eliminate the risk of punitive charges related to:

F_Pbl_21.1 Management of IPR issues, including limits on the use of submitted material by reviewers only for the purposes of the review;

F_Pbl_21.2 Ethical treatment of data related to research subjects (human subjects, animals, etc.);

[13] For example, up to two days of distance, on-line training.

[14] Ideally, free hosting. That may depend on the selected business model and the type of consequent co-operation between OpenScienceLink partners.

F_Pbl_21.3 Potential negative effects that could be claimed against the publisher of research data [15];

F_Pbl_22 Minimise the risks of unauthorised use /re-use/ of the journal concept, model, data, and related publications;

F_Pbl_23 Efficient management of the transfer of rights to re-use data and publications while protecting of the publisher's commercial interests.

Rationale for Goal 2: Even if sustainability is achieved, the publication of one or more open access data journals may not bring the profits expected by a commercial venture, in particular in the first years after launching a new journal. Therefore, a publishing company would expect to compensate the lack of profit by seeking to strengthen its competitive advantages overall.

Thus, a publisher [16] would expect that the publication of one or more data journals contributes to the company's competitive advantages and the OpenScienceLink platform and ecosystem should meet the following requirements:

F_Pbl_24 Increase the visibility of the publishing company, facilitate its presence in relevant professional and social networks and provide new networking opportunities;

F_Pbl_25 Link data journals to other relevant publications and company products and services;

F_Pbl_26 Provide access to know-how and awareness of innovation opportunities;

F_Pbl_27 Create opportunities for participation in funded national or international research and development, demonstration and innovation projects.

Overall, the publisher needs to have a clear view of their journals' positioning among the other journals in its field as well as monitor its progress in terms of a variety of perspectives, such as quality of published work, importance of authors in their domain, impact and potential. Hence, a series of requirements is set on the OpenScienceLink platform, including:

F_Pbl_28 View the evaluation of the journal(s) the publisher is responsible for as well of individual journal issues, published papers and data sets.

F_Pbl_28.1 Choose among the different metrics which are to be applied for the evaluation of the journal(s) that the publisher is responsible for.

F_Pbl_29 View the journal's ranking among the ones in the same field of research.

F_Pbl_29.1 Choose among the different metrics which are to be applied for the ranking of the journal(s) the publisher is responsible for.

F_Pbl_30 View and follow the evaluation of the author(s) who have published in the journal(s) that the publisher is responsible for

F_Pbl_30.1 Choose among the different metrics which are to be applied for the evaluation of the author(s) who have published in the journal(s) that the publisher is responsible for.

---

15 See also the previous requirement regarding complaints.

16 This is particularly valid for companies that have publishing of scientific literature as one of their lines business. This is the case with the publisher in the OpenScienceLink consortium – it combines research and consultations with scientific publishing. See www.procon.bg/business-areas.

2.2.4 Funders Requirements

Among the 'funders,' or "Funding Agencies," we considered corporate sponsors, venture capitals/funds and government funding agencies, as well inter- or supra-governmental organisations/ agencies, such as the EC.

Research councils and other funding institutions use peer review as a main tool in distributing funding. However, surveys performed about funding organisations showed that funders are challenged by an increasing number of applications and frequent *difficulty in finding willing and available reviewers*. A survey showed that most reviewers are motivated by a sense of professional duty and fairness despite not receiving academic recognition and frequently having to undertake this time-consuming work in their own time. Reviewers also indicated that they often receive requests for *reviews outside of their expertise* suggesting that funding organisations are having *difficulty targeting the appropriate reviewers.* Other problems that funding organisations face during the review process relate to poor quality reviews, receiving reviewers' reports with delays, and non-disclosure of potential conflicts of interest by reviewers. These findings are similar to the problems journals experience with peer review.

Several assessment criteria are used during peer review of grant applications:

Scientific quality of the research project

- importance to the international scientific community in the field(s) concerned;
- innovative aspects, i.e. the extent to which the project could break new ground scientifically;
- importance of the expected results for the discipline;
- clarity of the goals (hypotheses);
- appropriateness of the methods;
- quality of the cooperations.

Scientific quality of the scholars involved in the proposed project

- scientific *quality* and/or *potential* of the scientists involved;
- experience in organization and direction of national/international projects;
- expected importance of the project for the career development of the participants.

Additional aspects

- Can the project be expected to have implications for other branches of science?
- Is the project expected to have implications that go beyond academic purposes (potential industrial applications, results of relevance to society, etc.)?

The OpenScienceLink platform could be of significant use for funding agencies by providing a number of tools, including:

- Research trends analysis to investigate the potential impact of the proposed research;
- Novel metrics to evaluate the degree of novelty of the proposed research;
- Novel metrics to evaluate scientific quality and excellence of the of the scientists involved;

- Novel metrics to evaluate the expertise of the researchers in the proposed methodology;
- Novel tools to evaluate the quality of the institutions that form the consortium;
- Novel peer criterium describing the experience in organization and direction of scientific projects versus the age of the researcher.
- Novel tools to avoid bias.

For all those that fund research, it is the products of the grant funding that matter. Thus, evaluating these products may provide significant insight into successful and unsuccessful funding strategies and may provoke changes in the way that grants are allocated. In fact, the OpenScienceLink platform aims to provide several novel tools for evaluation of the research products.

In addition to facilitating the process of reviewing research proposals, OpenScienceLink is expected to be of utility for funding agencies by its research evaluation services.

Funding agencies, institutions that employ scientists, and scientists themselves, all have a desire, and need, to assess the quality and impact of scientific outputs. It is thus imperative that scientific output is measured accurately and evaluated wisely. Thus, there is a pressing need to improve the ways in which the output of scientific research is evaluated and all the parties involved are encouraging improved practices and methods in research assessment. Such steps are beginning to increase the momentum toward more sophisticated and meaningful approaches to research evaluation.

Outputs other than research articles (such as datasets) will grow in importance in assessing research effectiveness in the future, but the peer-reviewed research paper will remain a central research output that informs research assessment (San Francisco Declaration, 2012).

Rating the quality and strength of scientific output is an important part of the process of evaluating researchers and institutions, taking decisions about funding allocation and research direction and even deciding on the evidence-based health practices and policies. Scientists often arrive at their own judgments about the soundness of research output or technology assessments. Such judgments may differ considerably in the sophistication and lack of bias with which they were made (Lohr, 2004). Tools that meet acceptable scientific standards can facilitate these grading and rating steps and should be based on

- Quality measures: An expanded view holds that quality concerns the extent to which a study's design, conduct, and analysis have minimized biases in selecting subjects and measuring both outcomes and differences in the study groups other than the factors being studied that might influence the results;
- Quantity measures;
- Consistency measures.

The main objective of the OpenScienceLink pilot on "Scientific field-aware, Productivity- and Impact-oriented Enhanced Research Evaluation Services" is to introduce, produce and track new objective metrics of research and scientific performance, beyond conventional metrics, associated with conventional indices and impact factors. This new type of sciento-metrics enables research sponsors, funding authorities and governmental agencies to shape their research strategies, researchers to be evaluated based on a multitude of factors representative of their productivity, impact and domain rather than through simplified means such as the number of citations within a time period, important research work, in terms of potential, to be brought forward, among others.

F_Fnd_1 Present analysis of research trends related to the proposed research for assessing its novelty and expected impact;

F_Fnd_2 Present the evaluation of the applicants (researchers) based on a variety of metrics encapsulating aspects such as expertise and research potential;

F_Fnd_3 Present the evaluation of the proposed research work based on a variety of metrics encapsulating aspects such as novelty, expected impact, scientific excellence;

F_Fnd_4 Present the evaluation of the research team(s) applying for funding of their research work in terms of novelty, expected impact, scientific excellence;

F_Fnd_5 Present the evaluation of the organisations (research institutions, universities, academic schools, etc) who will participate in the research work applying for funding (Consortium members);

F_Fnd_6 Allow for the selection of the specific evaluation metrics to be used in each one of the above cases.

F_Fnd_7 Provide a list of experts whose research work is highly related to the one under application for funding.

F_Fnd_7.1 For each one of these experts, present his/her evaluation based on a variety of metrics encapsulating aspects such as expertise, impact of scientific work, and research potential among others.

F_Fnd_7.2 Allow for the selection of the metrics to be used for the experts evaluation.

F_Fnd_7.3 Allow for the ranking of the experts based on the evaluation metrics.

F_Fnd_8 Present emerging topics (or fields) and their trendiness in specified fields of interest for facilitating the selection of scientific topics to be included in research calls.

2.2.5 Requirements of Journalists and Press

This section details the requirements of journalists, press and mass media. By extrapolation the examination herein is expected to capture the requirements of wider society.

Previous research indicated that it is crucial that information for journalists, press and representatives of other media channels is free of charge and easily accessible. In order to improve communication between journalists and scientists (which is essential in communicating scientific information effectively) free access for journalists to scholarly literature databases and official sources of information—experts or institutions—is essential (Schwitzer, 2008; Kirby, 2011; Gilbers & Ovadia, 2011).

To this end, the OpenScienceLink webpage should have a separate platform for journalists, press, representatives of other media or bloggers that would include information pertinent to this group of stakeholders. Information blocs for the media in the dataset could be organized by:

- *News*: Most recent research/ scientific articles presented as press releases with a hyperlink to originals;
- *Experts:* Lists and profiles of experts available to comment and explain different scientific aspects covered by a media representative;
- *Journalists' Dataset:* a list of science journalists with personal profiles on the OpenScienceLink platform.

Besides, it is crucial that journalists would have access to all original data sources (original papers, comments by reviewers, editors, datasets, etc.) and research trends detected by OpenScienceLink platform.

Press Releases

Press releases would be the most suitable form to communicate scientific news (Ladle, et al., 2005). Short press releases should be generated for each scientific article and stored in the

dataset and accessible through the platform dedicated to journalists and media. Press releases in this platform should be organized in different categories based on the subject area (i.e., cardiology, dermatology, endocrinology, etc.). Upon pressing on selected category, an user should be able to access all press releases under that category.

Example:

First page >>	**Next page >>**	**Next page**
News *please select subject area* ... Cardiology Dermatology Endocrinology ...	**Press releases on cardiology** **July 5, 2013:** Study show that sleep apnea increases risk of sudden cardiac death **July 3, 2013**: Study: less than one in three US patients have BP **July 2, 2013**: New guidelines for management of heart failure ...	PRESS RELEASE **New guidelines for management of heart failure** July 2, 2013 Corresponding author and institution First paragraph Second paragraph Third paragraph

Recommended structure for press releases

Upon clicking on the title of the press release, a journalist should be able to access:

- Date of press release;
- Corresponding author's affiliation and contact information;
- First paragraph: short lead (3-5 sentences) indicating main ideas of the research with the answers to 5 questions (*What* have been done / found? *Why* the results are important? *Where* and *When* the research was conducted? *Who* performed the research / who participated?)
- Second paragraph: quotation of the leading authors, stressing the most important aspects of the research and explaining information in the leading paragraph.
- Third paragraph: additional information about the research, scholars or institution.

Managing lists of press releases

Users should be able to rearrange list of press releases in each category by:

1. Date and time of publication;
2. Number of views per original paper and press release;
3. Number of citations (Google scholar, etc.);
4. Number of mentions in popular press.

Dissemination of press releases

Once the press release has been placed in the dataset, information about it could be disseminated as follows:

ASAP	• *Social media:* OSL account should be created on Facebook, Twitter and other social media in order to share newest scientific

	information with journalists, media, as well as general publics, who follow OSL on social media. Once press release is put into the dataset, information about it should be sent via social media ASAP. • *News agencies:* collaboration with widely used news agencies by sending them prepared press releases.
Once a week	• *Newsletter:* users should be able to register for newsletter, which would be generated and send every week. Also, newsletters should be send to journalists and media which is specializing in particular scientific area (for this purpose dataset of media contacts should be created in advance). • *Blog:* OSL should have a separate blog where most interesting and important research would be discussed. Articles for the blog could be written by scientists themselves as well as by communication officer of OSL. At least one article should be published each week to keep blog active.

Who is responsible for communicating with the press and journalists?

To keep information in the press release as accurate as possible, scholars themselves would be responsible for creating a press release. This could be done by completing a short form.

Example:

Headline	……
Date	Day …… Month …… Year ……
Information about corresponding author and institution	• Author's full name: …… • Institution: …… • E-mail address: …… • Phone number: ……
First paragraph (lead)	…… ……
Second paragraph (quotation)	…… ……
Third paragraph (additional information)	…… ……

There should be clear indications for the scholars on how to create good headline, how to write a lead, quotation or what additional information is.

Example:

Recommendations for the headline	Headline should be short and clear. Difficult terminology should be avoided.
Date:	Generated automatically.
Information about corresponding author and institution:	Generated automatically with hyperlinks to the individual profile of the author and institution.

Instructions for the first paragraph:	In lay language explain *what* are the major *findings* of the research; indicate *why* they are important; *where*, *when*, and by *whom* the study was performed. *Do not* include methodology or theoretical outcomes, unless they are of great importance.
Instructions for the second paragraph:	Express your personal position about the research. Your quote should be related to the information in the first paragraph and support it.
Instructions for the third paragraph:	Provide additional information relevant to the study. For instance, methodological or theoretical background of the study, previous similar observations and findings, information about the project or institution, etc.

It is very important that OpenScienceLink personnel (communication officers) would be trained to review prepared press releases for final approval and pubic release.

Experts' Dataset

In case journalists need an expert to comment on specific topic, they can browse experts database and find the one. Experts in the dataset should be listed by subject area (e.g., experts in cardiology, experts in dermatology, experts in endocrinology, etc.). Once selected desired subject area, user should be able to see the list of experts with their full names, affiliations and fields of interests. Users should be able to manage list in:

- Alphabetical order of experts;
- Alphabetical order of institutions;
- Alphabetical order of fields of interest.

Also, users should be able to:

- search for certain expert by his name or surname
- select from the list of institutions (in order to see all the experts from selected institution)
- select from the list of fields of interest (in order to see all the experts interested in a selected field)

Example:

First page >>	**Next page >>**	**Next page**
Experts	**Experts on cardiology**	Surname, Name
please select subject area	…	Scientific degree
	Surname, Name, Scientific degree, Institution, Fields of interests	Institution
…		Fields of interest
Cardiology	Surname, Name, Scientific degree, Institution, Fields of interests	Contact information
Dermatology		Short Biography
Endocrinology	Surname, Name, Scientific degree, Institution, Fields of interests	
…	…	Publications

Once expert selected, user should be able to see his/her personal profile.

Experts' profiles

The major purpose of experts' profile is to provide journalists with general information about experts who would be able to provide comments and explain certain scientific information. An

expert's profile should provide general information about the scientist including his/her full name, scientific degree, institutional affiliation, fields of interest, contact information, and optionally contact data such as e-mail and phone number. In addition, a short biography and list of major publications with hyperlinks to their copies if possible should be added. Optionally, the expert's photo could also be added to the user's profile.

All scientists who are part of OpenScienLink would be asked to create personal profiles.

Journalists' Profiles

Each journalist should be able to create his/her own personal profile on the OpenScienceLink platform accessible with username and password. The profile would include:

- Journalist's name;
- Affiliation of media organization;
- Contact information including e-mail address, links to personal social media profiles (to facilitate communication between journalists and scientists or other journalists);
- Areas of interest, list of previous publications (based on that journalist would be able to receive weekly newsletters)

The OpenScienceLink platform should in addition be able to detect semantically-inferred relationships between the journalist and other OpenScienceLink stakeholders, including journalists, researchers, editors, funding agencies, etc.

F_Prs_1 Provide general information about experts who would be able to provide comments about a specific topic and explain certain scientific information;

F_Prs_1.1 The profile of the expert should include his/her full name, scientific degree, institutional affiliation, fields of interest, contact information, and optionally contact data such as e-mail and phone number as well as a photo;

F_Prs_1.2 The expert's profile should also include a list of major publications, with hyperlinks to their copies if possible.

F_Prs_2 Maintain and provide access to general information about experts who would be able to provide comments about a specific topic and explain certain scientific information;

F_Prs_3 View the top experts in a specific field of interest based on a series of evaluation metrics;

F_Prs_3.1 Rank the top experts in a specific field of interest based on a series of evaluation metrics;

F_Prs_3.2 Choose the evaluation metrics to be applied for extracting the top experts in a specific field of interest;

F_Prs_4 View the top research and academic institutions as well as research communities in a specific scientific field;

F_Prs_4.1 Rank the top research and academic institutions as well as research communities in a specific field of interest based on a series of evaluation metrics;

F_Prs_4.2 Choose the evaluation metrics to be applied for extracting the top research and academic institutions as well as research communities in a specific field of interest;

F_Prs_5 View the emerging research topics in a specific scientific area;

Experts in the dataset should be listed by subject area (e.g., experts in cardiology, experts in dermatology, experts in endocrinology, etc.). Once selected desired subject area, user should be able to see the list of experts with their full names, affiliations and fields of interests. Users should be able to manage list in:

- Alphabetical order of experts;
- Alphabetical order of institutions;
- Alphabetical order of fields of interest.

Also, users should be able to:

- search for certain expert by his name or surname
- select from the list of institutions (in order to see all the experts from selected institution)
- select from the list of fields of interest (in order to see all the experts interested in a selected field)

2.3 Requirements per OpenScienceLink Pilot

In the following sections, the aforedescribed requirements are grouped on a pilot basis, and are assigned with a priority level (core/essential/desired) which indicates the priority of consideration during the design and development phase of the platform. If possible, however, all of the requirements will be taken into account, but the core requirements will be given the absolutely highest priority.

2.3.1 Pilot 1: Research Dynamics-aware Open Access Data Journals Development

2.3.1.1 Purpose

Biomedical Dataset Journal will be an international, interdisciplinary journal for the publication of articles on original research datasets mainly in the scientific fields of cardiovascular research, endocrinology research, behavioral medicine and quality of life. The editor will encourage the submission of both ***clinical datasets*** as well as ***basic research datasets***. The journal aims to publish high quality, complete datasets in these areas which in the environment of the OSL platform will be able to be exploited and capitalized for the benefit of biomedical sciences.

The journal will maintain sections for regular articles, commentaries, as well as articles which could derive from the reuse of the published datasets. The journal will also publish review articles. Special issues may be planned aiming at the collection of datasets in specific areas of interest, such as for example datasets from patients with heart failure.

The publication of large, complete and high-quality datasets is considered to be very important for all researchers and departments in order to prove their significant ***experience and expertise in a scientific field***. Thus, publication in Biomedical dataset journal aims to become an important reference point for all researchers who aim to become established in a certain field and could be also used as an ***evaluation method for institutions and funding bodies***. A journal serving this purpose does not currently exist.

There are no universal well-structured repositories of scientific and research data for experimentation and benchmarking of pertinent research works in a given thematic area. Moreover, data journals are in several cases poorly linked to their relating scientif ic publications. As a result of the lack of consistent access to experimental datasets for benchmarking, evaluation and further research conduction upon them, the tasks of asserting

research experiments and verifying research results can be extremely time-consuming and in several cases the results are inaccurate.

The creation of a repository of well-structured and semantically linked datasets will permit the combined or comparative analysis leading to rapid evaluation of new hypotheses or

Regular articles of datasets must include a detailed description of the methodology used to obtain the submitted data in order to confirm the high quality of the data. Any interpretation and comments on data are welcome but not a prerequisite for regular articles.

Articles based on combination of previously published datasets should clearly refer to the datasets used, the hypothesis tested, the methodological and statistical approaches, the results and conclusions which provide novel insights into the field of biomedical sciences.

Review articles may compare methods or relative merits of datasets, the fitness of individual methods or datasets for specific purposes or how combinations might be used as more complex methods or reference data collections.

Decision for acceptance of submitted articles will be based on a peer-review process. Reviewers should have a significant experience in the specified field of research. Reviewer's decision should be based on:

1. The detailed description of the methodology used for collection of data according to standard guidelines. This will assure the high quality of the dataset.
2. The completeness of the provided information as compared to the need for others to reuse this dataset, or integrate it with other data.
3. Careful evaluation of newly described methods.
4. The completeness of the provided dataset in terms of number of measured variables in order to describe a population. Missing values should be kept to a minimum.
5. The number of subjects included in the submitted dataset. Datasets including small number of subjects in relation to the associated scientific field will not be accepted.

2.3.1.2 Key Stakeholders involved

Stakeholder	Description
Researchers	Submit Datasets to the platform for publication. Are potential evaluators.
Evaluators	Review submitted datasets.
Publisher/Editor/Funding Agencies	Find researchers that may act as potential evaluators (reviewers/referees) for submitted datasets. Organize and publish submitted datasets that are reviewed.

Table 2: Pilot 1 key stakeholders.

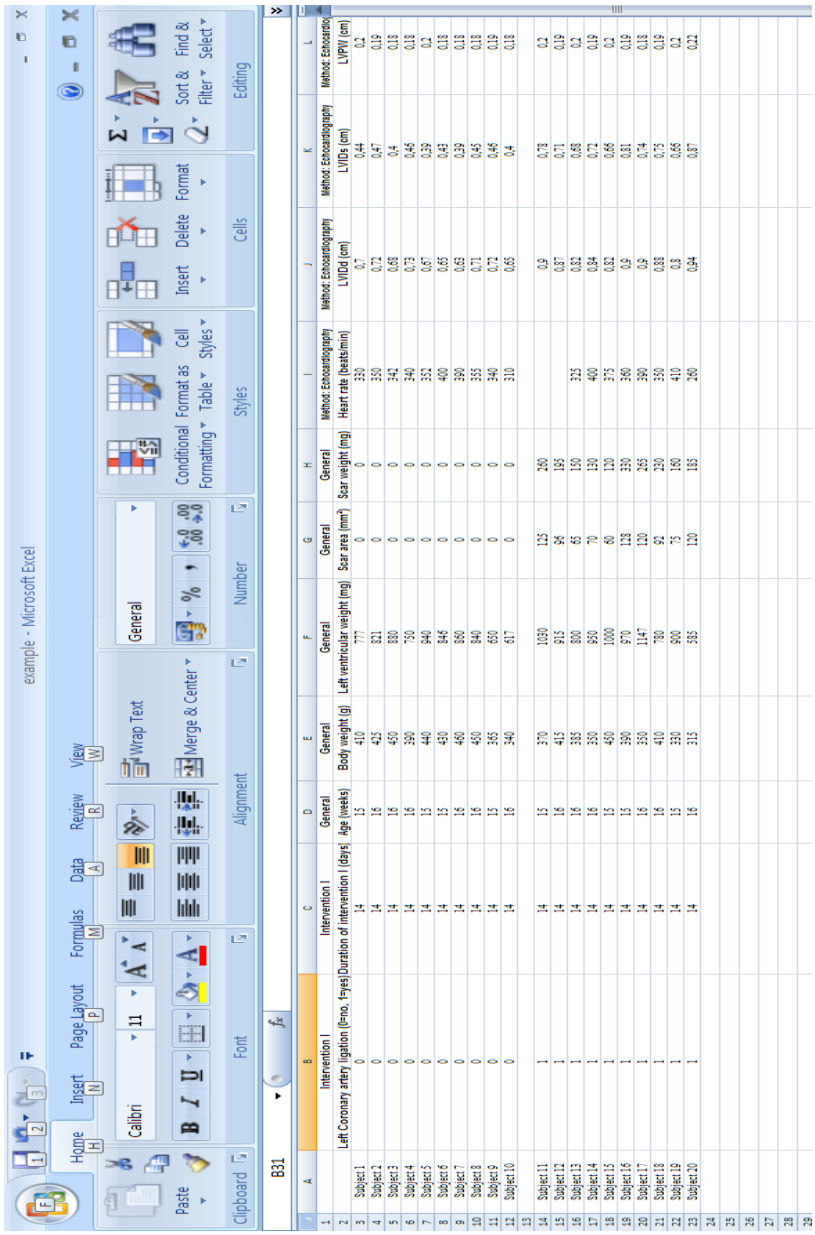
2.3.1.3 Overall Description

Before datasets submission authors need to be register with the OpenScienceLink platform. This process requires author details and also a set of keywords. Datasets are allowed to have any format as long as this is described analytically, with detailed instructions on how the submitted dataset should be read, interpreted and re-used in future works.

After registering to the OSL platform, the author will select ***the scientific field(s) of interest*** and whether the submission pertains ***an experimental or a clinical dataset***. Based on this input,

the platform will offer a number of choices in order to define the methodology used (e.g. echocardiographic evaluation, Cardiac magnetic resonance, Exercise testing, Blood measurements etc) and the variables measured in each case (e.g. for echocardiographic evaluation, Variable 1: Left ventricular internal diameter at diastole (LVIDd), Variable 2: Left ventricular internal diameter at systole (LVIDs), Variable 3: Ejection fraction (EF%) etc). This is a necessary step for building well-structured and semantically linked datasets. For this reason, different vocabularies are going to be created. In the case that a methodology and/or variable is not included in the build-in vocabularies, the author will have the choice to add new methodologies and/or variables as soon as he includes a clear definition. Figure 2 shows an example of a clinical dataset with anonymized subjects, and how this could be prepared for submission to the OpenScienceLink platform. Additional general remarks with regards to the preparation of clinical datasets for submission to the OpenScienceLink platform follow:

- Before submission authors need to prepare 1. a **word file** describing in detail the methodology used for collection of data. Papers should be clear, concise, and written in English. 2. One or more **excel files** containing the submitted dataset(s)
- Authors are advised to submit ***representative pictures or video*** whenever possible in order to show how measurements were performed.
- Authors should include as detailed information as possible about their methodology, including exact measurements as well as specific names and models of any equipment that is used.
- The word file must also contain all ***definitions of the variables*** measured as well as the ***scientific units*** used in each case. Units of measurement should be presented simply and concisely using ***System International (SI) units***.
- In the case of clinical datasets, all variables that could reveal the identity of the human subject should be deleted (e.g. name, phone number, date of birth, insurance number etc.)
- Variables in the datasets should be grouped according to the methodogy used to collect the data. e.g. Method: Echocardiographic evaluation, Variable 1: LVIDd, Variable 2: LVIDs, Variable 3: EF% etc
- ***Interventions*** should be described in datasets with numerical codes in order to serve the purpose of reuse of data and analysis e.g. Intervention I: coronary artery ligation, 0=no, 1=yes. The duration of intervention should always be defined as a separate variable e.g. Duration of Intervention I, 14 days etc,
- ***Categorical variables*** should also be described in datasets with numerical codes in order to serve the purpose of reuse of data and analysis e.g., Gender, 0=male, 1=female; or Killip classification, 0=I, 1=II, 2=III, 3=IV
- Drug administration can be defined either as an intervention or a categorical variable, e.g. Drug: Amiodarone, 0=No, 1=Yes. Duration of administration must be defined in a separate column.
- In the case of sequential measurements at different time-points in the same subjects, the time intervals should be clearly defined in the dataset e.g. Method: Echocardiographic evaluation at 2 weeks, Method: Echocardiographic evaluation at 8 weeks etc



	Intervention I		Intervention I	General	General	General	General	General	Method: Echocardiography	Method: Echocardiography	Method: Echocardiography	Method: Echocardio
	Left Coronary artery ligation (0=no, 1=yes)		Duration of intervention I (days)	Age (weeks)	Body weight (g)	Left ventricular weight (mg)	Scar area (mm²)	Scar weight (mg)	Heart rate (beats/min)	LVIDd (cm)	LVIDs (cm)	LVPW (cm)
Subject 1	0		14	15	410	777	0	0	330	0,7	0,44	0,2
Subject 2	0		14	16	425	821	0	0	350	0,72	0,47	0,19
Subject 3	0		14	16	450	880	0	0	342	0,68	0,4	0,18
Subject 4	0		14	16	390	750	0	0	340	0,73	0,45	0,18
Subject 5	0		14	15	440	940	0	0	352	0,67	0,39	0,2
Subject 6	0		14	15	430	846	0	0	400	0,65	0,43	0,18
Subject 7	0		14	16	450	860	0	0	390	0,63	0,39	0,18
Subject 8	0		14	16	450	840	0	0	355	0,71	0,45	0,18
Subject 9	0		14	15	365	650	0	0	340	0,72	0,45	0,19
Subject 10	0		14	16	340	617	0	0	310	0,65	0,4	0,18
Subject 11	1		14	15	370	1080	125	240		0,9	0,78	0,2
Subject 12	1		14	16	415	915	96	185		0,87	0,71	0,19
Subject 13	1		14	16	385	800	65	150	325	0,82	0,68	0,2
Subject 14	1		14	16	350	950	70	130	400	0,84	0,72	0,19
Subject 15	1		14	15	450	1000	60	120	375	0,82	0,66	0,2
Subject 16	1		14	15	390	970	128	330	360	0,9	0,81	0,19
Subject 17	1		14	16	350	1147	120	265	390	0,9	0,74	0,18
Subject 18	1		14	16	410	780	92	250	350	0,88	0,75	0,19
Subject 19	1		14	15	330	900	75	180	410	0,8	0,66	0,2
Subject 20	1		14	16	315	585	120	185	260	0,94	0,87	0,22

Figure 2: An example of a clinical dataset with anonymized subjects.

With regards to detailed genetic datasets, an example is given in Figure 3 on how these datasets could be structured for submission to the OpenScienceLink platform. Here, the first column describes an identification number of each sample. The columns 2 to 19 identify the alleles found for each of the loci typed. Columns 2 and 3 identify the two alleles for the microsatellite loci OR1. Each column is one allele present in this locus. Columns 4 and 5 are the alleles for the microsatellite loci OR3. Columns 6 and 7 are the alleles for the microsatellite loci CC1G02. Columns 8 and 9 are the alleles for the microsatellite loci CC1G03. Columns 10 and 11 are the alleles for the locus Rag1. Columns 12 and 13 are the alleles for the locus Cmos. Columns 14 and 15 are the alleles for the locus Rag2. Columns 16 and 17 are the alleles for the locus R35. Columns 18 and 19 are the alleles for the locus BDNF.

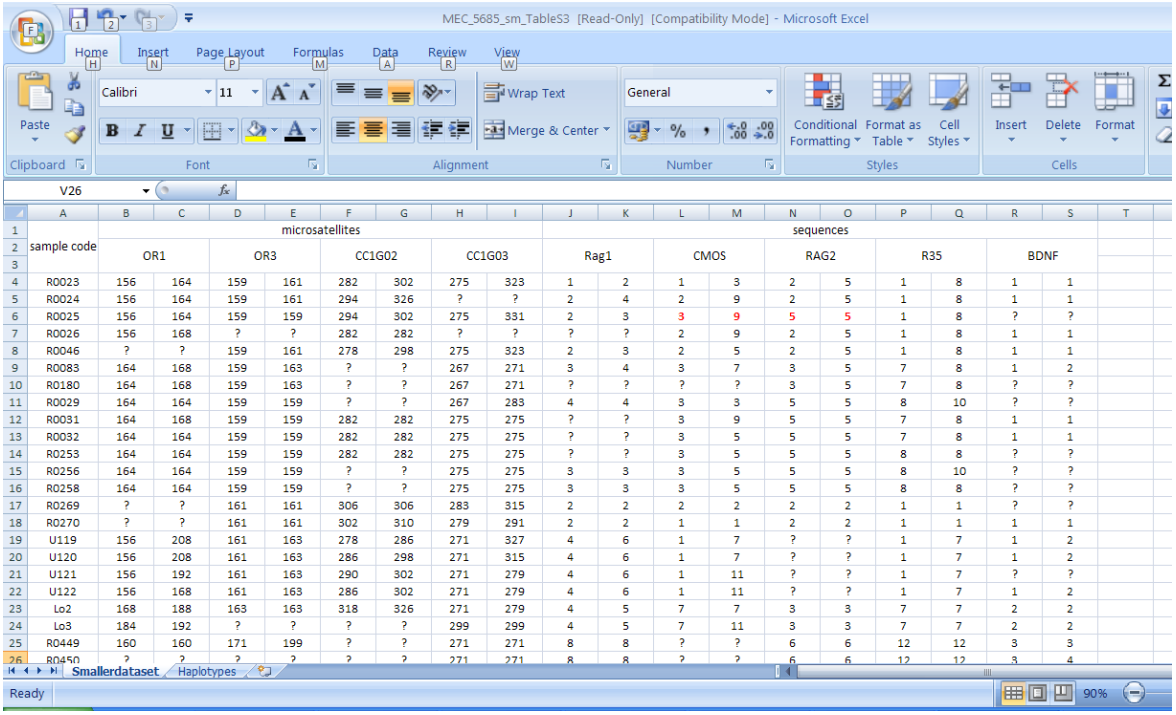


Figure 3: An example of a detailed genetic dataset.

A separate dataset should describe the GenBank accession numbers for each sample in this case. This is shown as an example in Figure 4. Each gene could be represented by two columns, corresponding to the two alleles found.

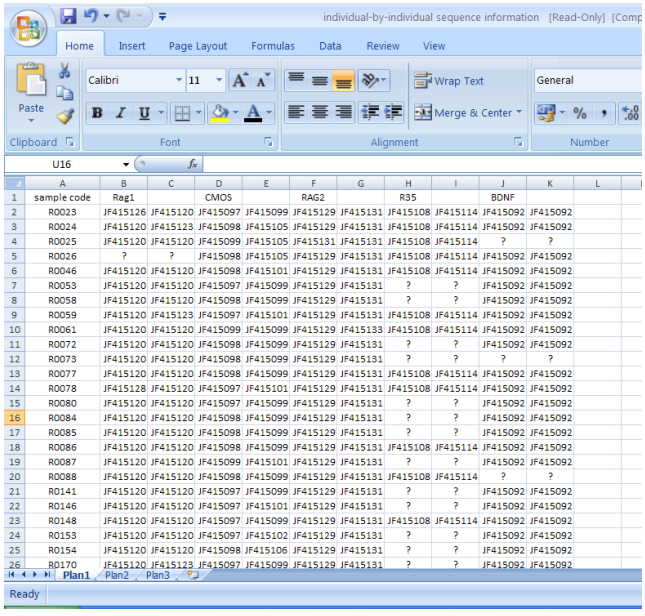


Figure 4: Example of GenBank accession numbers for each sample.

With regards to datasets that describe gene sequences, an example is shown in Figure 5. The number of aligned sequences will be identified as haplotype number followed by the GenBank reference number.

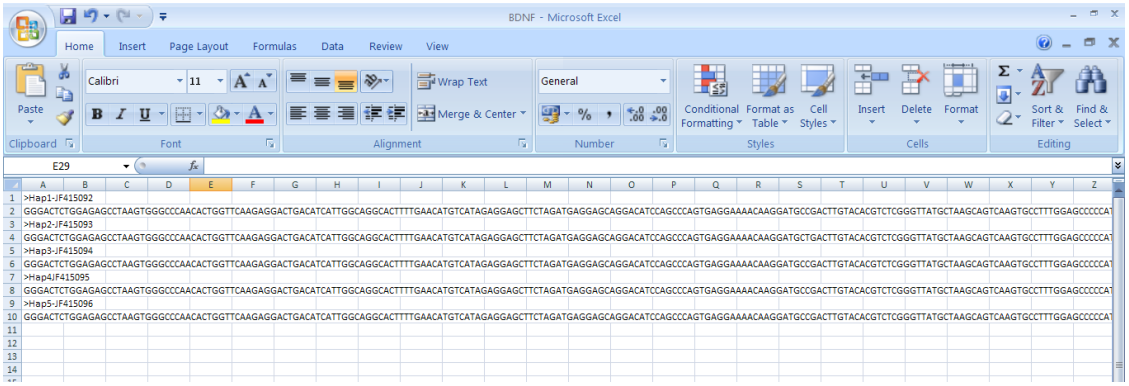


Figure 5: Example of a gene sequence dataset.

2.3.1.4 Requirements

Requirement ID	Description	Priority [core/ essential/ desired] [17]	Related Stakeholder Requirement ID
F_PR1.1	The researchers should be able to register and create a profile based on their academic/research interests.	Core	F_Rsr.1
F_PR1.2	The registered researchers should be able to comment/discuss previously published research works or datasets.	Essential	F_Rsr.6
F_PR1.3	The registered researchers should be able to upload their scientific data for sharing and publication under an open-access data journal.	Core	F_Rsr.7
F_PR1.4	Researchers should be given the ability to peer review research works and openly discuss and express opinions on the reviews and the review results.	Desired	F_Rsr.11
F_PR1.5	For an uploaded dataset, retrieve the metadata and the data.	Core	F_Evl_1
F_PR1.6	Retrieve related bibliography covering the topics or the datasets	Core	F_Evl_3
F_PR1.7	Retrieve related datasets to suggest	Desired	F_Evl_6

[17] **Core**: Requirements without which the pilot will not be able to deliver its main objective.
Essential: Requirements for which a short-term work-around could be developed, but over the long run, the requirements have to be there
Desired: Requirements which enrich the offered pilot services but without which the pilot will operate finely.
The pilot must be delivered with all its core requirements and a good portion of essential ones represented, with the overall plan to implement the remaining essential requirements in the following iterations.

Requirement ID	Description	Priority [core/ essential/ desired] [17]	Related Stakeholder Requirement ID
	comparisons.		
F_PR1.8	Accept or reject an invitation for review.	Core	F_Evl_7
F_PR1.9	Maintain the reviews within the platform for further processing, e.g. discussion with colleague researchers.	Desired	F_Evl_9
F_PR1.10	Identify trends, emerging research fields, and fields in decline;	Core	F_Pbl_1
F_PR1.11	Identify and prioritize potential members of the Editorial Board vis-à-vis initial and emerging research fields to be covered by the journal;	Essential	F_Pbl_5
F_PR1.12	Identify potential contributors to the journal;	Core	F_Pbl_6
F_PR1.13	Identify and prioritize candidate reviewers per submitted contribution while eliminating potential conflict of interests;	Core	F_Pbl_7
F_PR1.14	Facilitate the review process.	Essential	F_Pbl_8
F_PR1.15	Track the performance of reviewers.	Desired	F_Pbl_9
F_PR1.16	Provide high visibility of the data journal.	Core	F_Pbl_10
F_PR1.17	Provide permanent access, quickly and in long-term.	Core	F_Pbl_11
F_PR1.18	Sustain the interest in the journal by creating and nurturing a community of stakeholders (including readers/users of journal publications).	Desired	F_Pbl_12
F_PR1.20	Produce and disseminate calls for papers and other announcements (via listservs, dedicated sites, professional networks, etc.);	Core	F_Pbl_14
F_PR1.21	Automate the publication management process.	Core	F_Pbl_15
F_PR1.22	Automate the maintenance process.	Desired	F_Pbl_16
F_PR1.24	Link data journals to other relevant publications and company products and services	Core	F_Pbl_25

Table 3: Pilot 1 requirements.

2.3.2 Pilot 2: Novel open, semantically-assisted peer review process

2.3.2.1 Purpose

The reviewing of manuscripts is an essential step in the publication process. This Pilot aims at offering a novel peer-review process methodology, which will help overcome a number of problems currently encountered by researchers, referees and editors.

The peer-review system faces two common criticisms: 1. that the system wrongly rejects scientifically valid papers, and 2. that the system wrongly accepts scientifically flawed papers. There are many examples where journals had to retract papers because errors, or even outright fraud, went undetected by the reviewers. Studies have found that peer-review has little effect on improving the quality of articles. Peer-review publication is time-consuming and expensive and often excludes people for no good reason.

The process of peer reviews often leads journals to make subjective decisions based on the comments of a few referees. Researchers complain that even if their works are scientifically sound, articles are rejected based and as a result, are often submitted to successive journals before being accepted and then published many months or even years after their first submission. In this way, the 'review burden' on the academic community is multiplied by the fresh review of each new re-submission. Based on a number of new tools *OpenSciencelink* has the potential to substantially improve the speed and efficiency of the review process, which in turn will accelerate research itself.

One of the main critisisms of the peer-review system is that it is not a reliable, objective and consistent process and a little better than you would expect if decisions were taken by chance. In fact, the opinion of a small number of individuals (usually 2 peer reviewers and the editor) determines whether a paper is published in a journal. If both advise publication the editor sends it to the publisher. If both advise against publication the editor rejects the paper. If the reviewers disagree the editor sends it to a third reviewer and does whatever he or she advises. This pastiche is little better than tossing a coin. By extension, the opinion of those people (who act on the paper in advance of publication, and without the benefit of evaluation over time by the wider community) exerts a disproportionate influence on what is read by the community.

All scientists are humans. Scientific dogmas, political concerns, economic factors, lab-rivalry, support for one's friends, and other normal human elements are never completely divorced from the peer-review process.

Peer review has been shown to be inefficient in detecting errors or fraud. BMJ has performed studies where major errors into papers were inserted and then sent to many reviewers. Nobody ever spotted all of the errors. Some reviewers did not spot any, and most reviewers spotted only about a quarter. These results clearly show that a peer-review process aiming to detect most errors or frauds cannot be based only on 2 or 3 referees, but has to rely on the critical capacity of more.

Relevant work by other scientists is not frequently cited in published papers. Moreover, a referee should call to the editor's attention any substantial similarity between the manuscript under consideration and any published paper or any manuscript submitted concurrently to another journal. The process of searching and finding these relevant publications is currently inefficient (produces a large portion of irrelevant results) and time-consuming.

Finding the appropriate and available reviewers is often one of the main problems of editors. Reviewers are not paid and do not often have enough motivation to participate in the review process. Some journals estimated that if the time spent reviewing was calculated as productive time (like doing original research) the cost of peer review per paper was of the order of £100 to £1000. Based on this, it is not a surprise that a high percentage of reviewers rejects the invitation and the editor often has to address to a "friend" or ask the authors to suggest potential reviewers. In addition, there is no current evaluation system for reviewers and even though participation in peer-review process has been considered by some institutions and bodies as a benchmarking tool, there is no way to easily find information and statistics about a researcher's experience in peer-review. Thus, a system for recording and evaluating the activity of reviewers would be expected to greatly increase their willingness to participate in the peer-review process.

Another issue has to do with the need to find contact information of the appropriate reviewers. Even if the editor is able to spot a reviewer with experience in the specific scientific field, he won't be able to contact him. Thus, most editors prefer to select reviewers from a list of

researchers that they have contact information. Furthermore, finding contact information of eligible reviewers through published articles is not so efficient since this information often refers only to the corresponding author.

Identifying conflict of interest of selected referees has been proved to be a major problem. A referee should not evaluate a manuscript authored or co-authored by a person with whom the referee has a personal or professional connection if the relationship would bias judgment of the manuscript. CoIs may be individual, institutional, financial, academic, or personal. Especially non-financial CoIs may involve rivalry among disciplines, cronyism, and geographic and academic biases. They are important because they can impact the fairness of the peer-review process. Editors do not have the reliable tools to identify these personal or professional connections between authors and reviewers. Things become even more complex by the fact that the suggestion of potential reviewers by the authors has become the main policy of several journals in order to speed up the process. Editors have just to rely on the referees' statement declaring that there is no conflict of interest. However, disclosure statements may also have limitations, as competing interests may be underreported or misreported particularly those of a non-financial nature.

In order to complete their task, reviewers have to follow a stepwise approach by answering a number of questions in relation to the submitted work. Being able to adequately answer these questions, they often have to search for information in the literature with tools that are inefficient and time-consuming. Some examples of the queries they have to answer are shown below:

1. Have the authors previously worked in this research field?
2. What is the authors' previous experience in the methods described in this manuscript? Is the manuscript technically sound?
3. Has the statistical analysis been performed appropriately? Does the manuscript need an evaluation by an expert statistician?
4. If this is an experimental study, did the authors treat animals according to ethical rules of animal experimentation?
5. If this is a clinical study, have the authors obtained informed consent of included patients?
6. If this is a treatment study (clinical trial), is this study registered? Does the manuscript refer to the trial registration number?
7. Do you believe that this work provides significant new insight in relation to previously published information in this scientific field?
8. Do you believe that this manuscript could be of general, specific or narrow interest to the scientific community?
9. Are conclusions supported by presented data?
10. Is scientific reasoning and argumentation sound?
11. Do the authors adequately refer and comment on relevant work published by other researchers?
12. Evaluate strength of evidence including ***internal validity*** (the extent to which studies yield valid information about the populations and settings in which they were conducted), ***external validity*** (the extent to which studies are relevant and can be generalized to broader patient populations of interest), and ***coherence or consistency*** (the extent to which the body of evidence makes sense, given the underlying model for the clinical situation).

2.3.2.2 Key Stakeholders involved

Stakeholder	Description
Researchers	Submit Datasets to the platform for publication. Are potential evaluators.
Evaluators	Find related information: topics, collaborators, colleagues, fundings, publications; to judge submissions.
Publisher/Editor/Funding Agencies	Find researchers that may act as potential evaluator (reviewers/referees) for certain topics.

Table 4: Pilot 2 key stakeholders.

2.3.2.3 Overall Description

OSL aims to make the review process more democratic, to ease the assignment of editors and reviewers to manuscripts, to add benchmarking in the contribution of reviewers in improving scientific articles, and moreover to shield editors and reviewers from conflicts of interests and allegations.

The platform will create ***a database of researchers*** that will include their profile (name, department, title, position, address, e-mail, main areas of scientific interest, areas of methodological expertise). The database will be created based on the registration of Opensciencelink users, on existing literature and published papers and social media. Information in this researchers database will be semantically linked with available ontologies. This semantically linked data will allow to create ***groups of experts*** related to 1) scientific fields, or 2) to methodological expertise area

The editor responsible for the peer review process will be able to log in and define the scientific fields, and the methodological expertise area related to the submitted manuscript. Then, OSL platform using the semantically linked data from the ***database of researchers*** will help to identify the most appropriate reviewers for evaluating each paper, from ***the group of experts*** scientifically related to the specific research topic, based on their research activity and output. The Platform will propose a list of potential reviewers and will also provide their contact details. The next step will be to filter out from the proposed list of reviewers (for a specific research work) the scientists who directly or indirectly relate to the authors of the work under review. The Platform performs this evaluation and notifies the editor in cases that conflict of interest is detected. The concept is that the platform will be able to provide a large number of potential reviewers for each scientific field and the editor will be able to contact immediately 10-20 selected expert reviewers.

Each reviewer will be able to log in the OpenScienceLink Platform in order to review the assigned work. The Platform aids the reviewer by providing ***automatically-generated results in a number of pre-defined queries***. First of all, the Platform provides a list of papers which are semantically related to the paper under review. These papers serve as state of the art work to which the paper should be compared in terms of similarity and advancement. This could help the reviewer decide whether the submitted work provides significant new insight in this scientific field. Additionally, the reviewer can suggest papers from the list for inclusion as references to the research work under review, given their semantic relevance to the research topic the work deals with. A complete list of previous publications of the submitting authors could be very useful to assess the experience in the specific scientific field. Moreover, a list of publications of the submitting authors concerning the methods described in the submitted work would be helpful in order to define the methodological expertise. Finally, by providing ***a research trend analysis*** about the specific research area, the OSL platform will assist the reviewer to evaluate whether this manuscript could be of general, specific or narrow interest to the scientific community. The reviewer can perform, at any time, semantic searches across a variety of open access data sources, in order to identify relevant literature, which will help him/her evaluate the paper or dataset under review.

The reviewer further uses the platform's semantic filtering mechanism, which enables the semantic filtering of the initial search results s/he is presented with. This filtering takes place

through the ontologies incorporated in the OpenScienceLink Platform. In fact, the ontologies serve as a dynamic table of contents for each result set with which the reviewers are presented.

The Platform allows the reviewer to apply a variety of ranking factors on the retrieved results, such as relevance, timeliness, quality of publications' fora, "authority" of author, etc., according to his/her preference.

For each ontologically annotated term in the presented results (either of a direct reviewer's search or of one of the automatic results of the above Services) the reviewer is presented with the respective data from the Linked Data sources, in order to aid the reviewer understand terms which s/he might not be familiar with.

The reviewers will prepare and submit their reports within a limited time frame.

The comments-criticisms of all reviewers will be unified in a list independent of the submitting reviewer. An ***on-line discussion forum for the specific reviewers group*** will be created and each reviewer will be able to vote in favor or against each comment-criticism that has been raised. If less than 50% of the reviewers vote in favor of a comment, then this comment will be rejected. Comments with more than 50% of the reviewers voting in favor will be accepted and sent to the authors. A detailed response prepared by the authors to each comment raised by the referees will be presented to the discussion forum and each reviewer will be asked to vote if he/she is satisfied with the authors answer. The decision of whether the manuscript will be accepted, rejected or granted revision, will be the decision of the editor based on the results of the discussion forum.

Another tool that this platform aims to create is a ***reviewers evaluation tool*** based on: 1. personal scientific output and scientometrics, 2. experience and performance of each scientist in the review process. This tool will provide significant motivation to researchers to participate in the peer-review process and have an objective estimation of their effort. Moreover, based on this reviewers' evaluation tool, the opinion of each reviewer could have a different impact on the review process. This means that the opinion of someone, whose scientific work and reviewer experience are highly regarded, carries more weight than the opinion of someone whose rates are poor. This could be a dynamic *process.*

Interactive open post-review discussion

In the Interactive open post-review discussion following the publication of a paper, the following types of interactive comments can be submitted via the OpenSciencelink platform:

Short Comments (SC) can be posted by any registered member of the scientific community (free online registration). Such comments are attributed, i.e. published under the name of the commentator.

Referee Comments (RC) can only be posted by the referees involved in the peer-review of the discussion paper. They can be anonymous or attributed (according to the referee's preference).

Editor Comments (EC) can only be posted by the Editor of the paper.

Author Comments (AC) can only be posted by the contact author of the paper on behalf of all co-authors.

Every registered member of the scientific community may publish Short Comments as defined above. The authors of a paper are automatically informed by e-mail about the publication of comments in the Interactive Post-review Discussion of their paper. The authors of the paper have the option (but no obligation) to reply by publishing their own Short Comments on behalf of all co-authors. Publication alert services will also be available to other members of the scientific community. The publication of interactive comments is supervised by the Editors, who have the option of censoring comments that are not of a substantial nature and of direct relevance to the issues raised in the discussion paper or which contain personal insults. The editorial board reserves the right to exclude abusive commentators.

The comments submitted during the open-identity post-review discussion are annotated with terms from the ontologies incorporated into the OpenScienceLink Platform. This way, the editor is able to easily navigate through the comments and filter them based on the ontology terms, and

have a more efficient and fast way of having a clear view on the scientific community's feedback on the submitted reviews. The Platform also helps the publisher to evaluate the confidence of the comments received during the open-identity post-review discussion, based on the researcher's relevance and impact in the field.

After the open discussion no more Short Comments and Referee Comments are accepted, and the Editor of the paper has the opportunity to publish final Author Comments and Editor Comments, respectively. The final response phase will have a definite time limitation (4 weeks to 8 weeks) and automatically terminated upon submission of a revised manuscript. Before submitting a revised version of their manuscript for publication (second stage of publication), the authors are supposed to have answered the Referee Comments and relevant Short Comments cumulatively or individually in one or more Author Comments.

Funding Agencies and peer-review process

Peer review is the main 'equipment' used by research councils and other institutions to distribute funding. Surveys performed about funding organisations showed that funders are challenged by an increasing number of applications and frequent ***difficulty in finding willing and available reviewers***. A survey showed that most reviewers are motivated by a sense of professional duty and fairness despite not receiving academic recognition and frequently having to undertake this time-consuming work in their own time. Reviewers also indicated that they are often sent requests for ***reviews outside of their expertise*** suggesting that funding organisations are having ***difficulty targeting the appropriate reviewers. Other problems that funding*** organisations ***face during the review process are: r***eceiving poor quality reviews, receiving late reviewers' reports, reviewers not declaring their conflicts of interest. These findings are similar to the problems journals experience with peer review.

Several assessment criteria are used during peer review of grant applications:

1. **Scientific quality of the project**
- importance to the international scientific community in the field(s) concerned
- extent to which the project could break new ground scientifically (innovative aspects)
- importance of the expected results for the discipline
- clarity of the goals (hypotheses)
- appropriateness of the methods
- quality of the cooperations

2. **Scientific quality of the scientists involved**
- scientific *quality* and/or *potential* of the scientists involved
- experience in organization and direction of national/international projects
- expected importance of the project for the career development of the participants

3. Appropriateness of the financial planning

4. Additional aspects
- Can the project be expected to have implications for other branches of science?
- Is the project expected to have implications that go beyond academic purposes (potential industrial applications, results of relevance to society etc.)?

It becomes obvious that new tools introduced by OSL platform could be of significant use for funding agencies. These tools may include:

- Research trends analysis to investigate the potential impact of the proposed research
- Novel metrics to evaluate the degree of novelty of the proposed research

- Novel metrics to evaluate scientific quality and excellence of the of the scientists involved
- Novel metrics to evaluate the expertise of the researchers in the proposed methodology
- Novel tools to evaluate the quality of the institutions that form the consortium
- Novel peer criterium describing the experience in organization and direction of scientific projects versus the age of the researcher.
- Novel tools to avoid bias.

For all those that fund research, it is the products of the grant funding that matter. Thus, evaluating these products may provide significant insight into successful and unsuccessful funding strategies and may provoke changes in the way that grants are given. In fact, OSL platform aims to provide several novel tools for evaluation of the research products.

2.3.2.4 Requirements

Requirement ID	Description	Priority [core/essential/desired]	Related Stakeholder Requirement ID
F_PR2.1	The user (evaluator, editor, publisher) has to login to authenticate.	Core	
F_PR2.2	The evaluator has to set up a profile with links to his/her articles. This will enable recruitment by editors and funders.	Core	F_Evl_14
F_PR2.3	The evaluator includes major topics s/he is working on into the profile.	Core	F_Evl_14
F_PR2.4	The researcher retrieves an invitation to be a reviewer.	Core	F_Evl_7
F_PR2.5	The user uses a form to accept or reject the invitation. In the later case s/he should put a recommendation of a colleague that could do the review.	Core	F_Evl_7
F_PR2.6	The evaluator submits a review to the platform for the authors / researchers, editor and funder to access.	Core	F_Evl_10
F_PR2.7	The evaluator uses a form to put the review results.	Desired	F_Evl_10
F_PR2.8	The publisher sets up a form in which the evaluator can place his/her review.	Desired	F_Pbl_12
F_PR2.9	The evaluator can submit a request to get related information for the uploaded article, including other publications of the author.	Core	F_Evl_17
F_PR2.10	The publisher submits a request to retrieve potential evaluators based on an article text. Also unregistered researchers (e.g. authors) are retrieved.	Core	F_Pbl_12, F_Pbl_15

Requirement ID	Description	Priority [core/essential/desired]	Related Stakeholder Requirement ID
F_PR2.11	The evaluator retrieves the article text via the platform.	Core	F_Evl_2
F_PR2.12	The evaluator can further filter by topics and other biomedical concepts, i.e. semantically search.	Core	F_Evl_3
F_PR2.13	The evaluator can retrieve a collaboration network for the target research area.	Core	F_Evl_5
F_PR2.14	The registered researchers should be able to comment/discuss previously published research works or datasets.	Desired	F_Rsr_6
F_PR2.15	Researchers should be given the ability to peer review research works and openly discuss and express opinions on the reviews and the review results.	Core	F_Rsr_11
F_PR2.16	For an uploaded dataset, retrieve the metadata and the data.	Core	F_Evl_1
F_PR2.17	Retrieve related bibliography covering the topics or the datasets	Core	F_Evl_3
F_PR2.18	Retrieve cited bibliography.	Essential	F_Evl_4
F_PR2.19	Retrieve related datasets to suggest comparisons.	Core	F_Evl_6
F_PR2.20	View reviews of other evaluators on already published work.	Desired	F_Evl_8
F_PR2.21	Maintain the reviews within the platform for further processing, e.g. discussion with colleague researchers.	Essential	F_Evl_9
F_PR2.22	Retrieve information on the scientific direction and trends of a specific topic.	Desired	F_Evl_11
F_PR2.23	Evaluators are anonymous w.r.t. the authors / researchers and colleague evaluators.	Desired	F_Evl_12
F_PR2.24	Evaluators provide a valid email address, so they can be contacted by editors and funders.	Core	F_Evl_15

Table 5: Pilot 2 requirements.

2.3.3 Pilot 3: Data mining for Biomedical and Clinical Research Trends Detection and Analysis

2.3.3.1 Purpose

To discover and analyse research trends in open access biomedical and clinical research texts. The identification and analysis of such trends is essential for the efficient allocation of research funding (by private sponsors and governmental agencies), for the overall planning of research strategies, for publishers to position their next journal issues and/or journals on scientifically hot areas of research and for researchers to have an overview of the research dynamics, among others.

2.3.3.2 Key Stakeholders involved

Stakeholder	Description
Researcher	Requests for current and/or rising trends in their scientific area for starting their research efforts or re-directing their existing ones Requests for research group and/or institution dynamics (growing, shrinking, etc) in order to choose one for collaboration with or for joining Requests for research community dynamics (growing, shrinking, etc) in order to choose one for collaboration with or for joining
Universities/Academic Institutes	Requests for trends in their scientific area for opening up new positions, announcing scholarships, re-directing funds for research, etc Requests for research dynamics of a researcher (new in area, promising, etc)
Publishers	Requests for rising trends in the scientific area of their journal for making go/no-go decisions about journals and boosting dissemination of existing ones. Request for validation of the research potential of specific topics for evaluating the need for stopping circulation of existing journals.
Enterprises with R&D departments (e.g., pharma)	Request for research topics and/or fields in their infancies which are expected to rise for prioritising their research efforts, identifying research areas for funding and resource allocation, and reaching earlier and more reliable go/no-go decisions about research proposals. Request for validation of the research potential of specific topics for early identification of failure and repositioning of resources to other research areas of great interest and with high expected impact.
Research Sponsors and Funding Authorities	Request for hot research topics in their infancies - in a scientific area - for focusing their funding opportunities and increasing the outcome potential, and, thus, achieving more efficient and effective allocation of governmental and private funds to research.
Media and Press	Request for trends in the overall domain for spotting areas for which they could publish timely, interesting and hot information about R&D developments Be presented with the top 20 hot trends, the top 20 scientists, the top 20 universities and research centers in a scientific area

Table 6: Pilot 3 key stakeholders.

2.3.3.3 Overall Description

This pilot emphasizes the ability of the OpenScienceLink Platform to discover and analyze research trends. The identification and analysis of such trends is essential for the allocation of research funding (by private sponsors and governmental agencies), as well as for the overall

planning of research strategies. The Pilot is primarily empowered by the advanced data mining capabilities of the Platform, on the basis of data for the biomedical and clinical research field. Yet it will be developed in a general manner, applicable to multiple domains.

Research trend analysis will be used to evaluate the degree of novelty of a publication or a research proposal. Furthermore, it is important to assess research trends in evaluating the impact of a research work in terms of citations. Thus, citation analysis should be weighted against research trends. Thus, the identification of scientific fields related to each publication will be the basis for these analysis. If for example, a user of the platform would like to compare research trends in 3 different fields of cardiovascular research, such as 1. "Ischemic preconditioning" and heart, 2. "Stem cells" and heart and 3. "thyroid hormone" and heart, he will be able to receive a graph with number of publications per year, as shown below, in Figure 6.

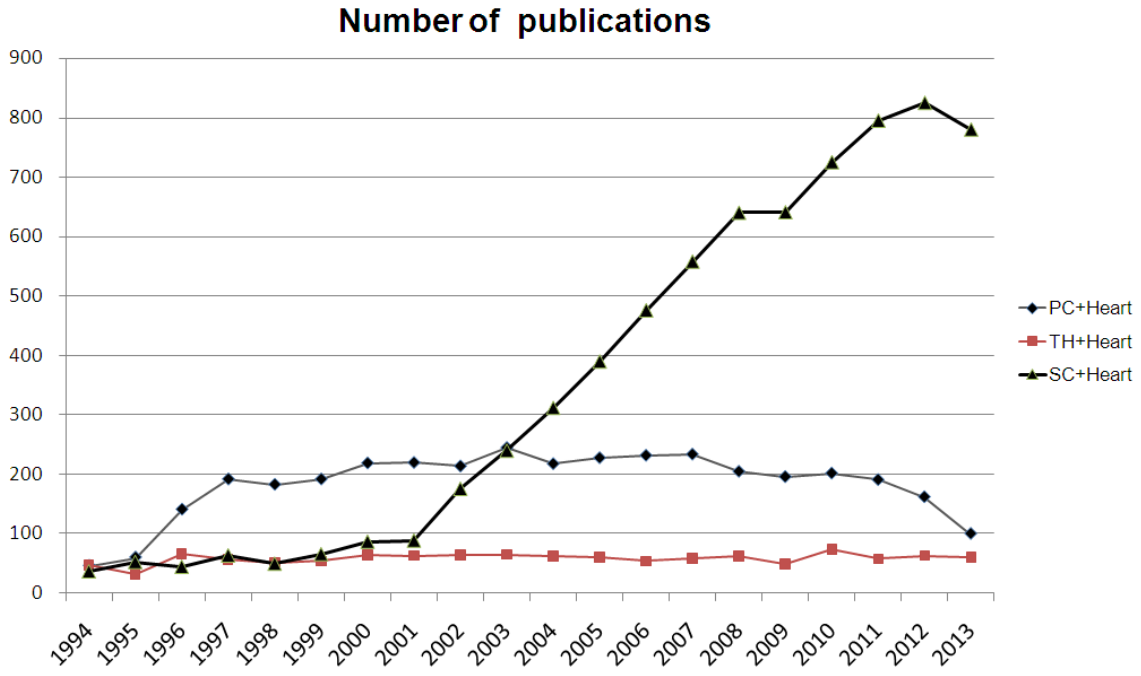


Figure 6: Example of trend analysis graph.

From this Research trend analysis graph a number of conclusions can be extracted:

The scientific field of "*PC and heart*" showed an expansion between 1995 and 1997, which resulted in a plateau from 1998 to 2010 and from 2010 to 2013 the field is in decline. Thus, this field has been a trend, but it shows signs of exhaustion.

The field of "*SC and heart*" was at a low steady-state level of publication from 1994 to 2000, but between 2001 and 2012 it showed a remarkable rate of expansion. Thus, this field is a current research trend.

The field of "*TH and heart*" is at a low steady-state level of publication from 1994 to 2013. Thus, this field has never become a research trend till now and probably has a high potential.

Besides the temporal distribution of publication, the geographical distribution could be another important component of research trend analysis. This data could be important for institutions, such as the NIH or the World Health Organization to take decisions. The aim of the OSL platform will be to provide automated generation of this information represented on a map showing the number of studies performed per city for a given scientific area. For example, a world map of studies performed in the field of "*iodine deficiency*" is shown below. From this geographical research trend analysis, it can be concluded that iodine deficiency has been extensively studied in Europe, while very few studies have been performed in Australia or Africa. Based on this evidence, WHO could decide to promote this area of research in these regions. Moreover, NIH may decide to give priority to funding in studies of iodine deficiency, since the number of studies

in this field is significantly lower in the USA as compared to the number of studies conducted in Europe.

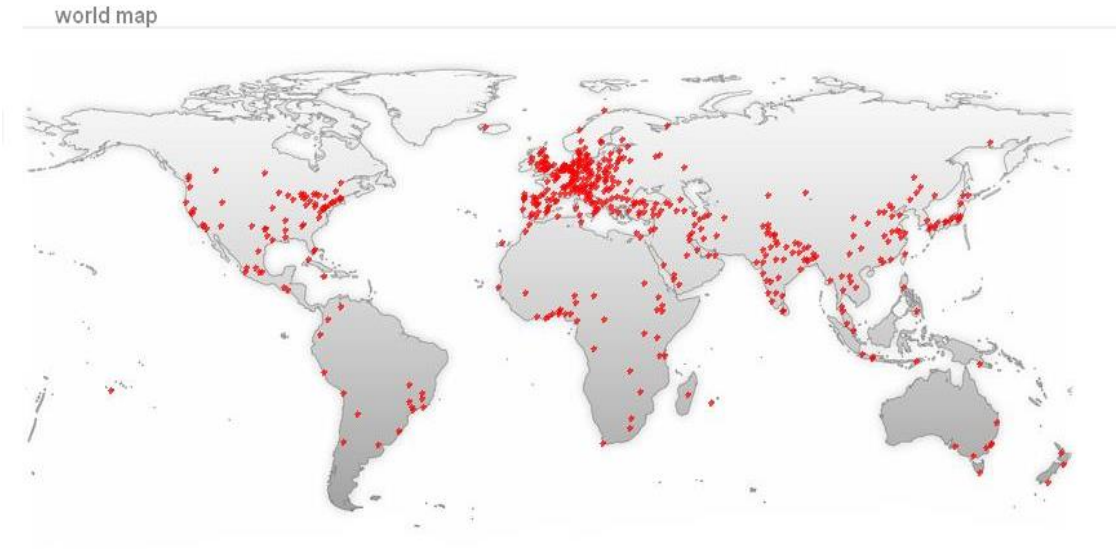


Figure 7: Example of world map of 'iodine deficiency' studies, as produced by GoPubMed.

2.3.3.4 Requirements

Requirement ID	Description	Priority [core/essential/desired]	Related Stakeholder Requirement ID
F_PR3.1	The registered users should be able to see the advances (emerging fields and trendiness) in their fields of interest.	Core	F_Rsr_3, F_Fnd_8, F_Prs_5
F_PR3.2	The registered researchers should be able to see the top authors, research institutes and countries in the field of their study and interest.	Essential	F_Rsr_4, F_Prs_3, F_Prs_4
F_PR3.3	The registered researchers should be able to receive notifications based on pre-selected criteria with regards to newly published research work or data sets in their fields of interest.	Desired	F_Rsr_14
F_PR3.4	Retrieve information on the scientific direction and trends of a specific topic.	Core	F_Evl_11, F_Fnd_1
F_PR3.5	Evaluators employed by funding agencies have access to already funded	Desired	F_Evl_16

Requirement ID	Description	Priority [core/essential/desired]	Related Stakeholder Requirement ID
	work in the respective fields.		
F_PR3.6	Identify trends, emerging research fields, and fields in decline;	Core	F_Pbl_1, F_Prs_5
F_PR3.7	Maintain awareness of competitive publications (both traditional and open access journals and conference proceedings);	Essential	F_Pbl_2
F_PR3.8	Provide access to know-how and awareness of innovation opportunities;	Desired	F_Pbl_26

Table 7: Pilot 3 requirements.

2.3.4 Pilot 4: Data mining for proactive formulation of scientific collaborations

2.3.4.1 Purpose

To allow for researchers to investigate potential, non-declared relationships with other researchers, research groups and/or existing communities, as well as to build an open, dynamic scientific community for a given scientific research topic.

2.3.4.2 Key Stakeholders involved

Stakeholder	Description
Researcher	Be suggested with potential collaborations with other researchers and/or research groups and communities Be presented with the opportunity to join dynamically formed research communities within their field of interest
Publisher/Editor	Be suggested with potential collaborations with researchers for leading an issue and/or journal

Table 8: Pilot 4 key stakeholders

2.3.4.3 Overall Description

This pilot focuses on automatically performing a series of intelligent matchings among researchers based on their dynamically generated research interests, as they are inferred through their published scientific work (including data sets, papers, articles, etc). The aim of these matchings is to detect and propose collaborations among researchers who have not worked together and have no obvious, detectable connection, although they share the same or common research interests. Hence, the aforementioned matchings are filtered by taking into consideration their past and existing collaborations as well as their deduced acquaintance in social networks. Similar matchings are made between researchers and research groups as well as communities based on their scientific topics in order to propose their collaboration and/or participation in them. For this purpose, the profile of the research group/community needs to be publicly available and accessible through the Internet or, alternatively, the members of the group/community along with their scientific work.
Publishers and editors also participate in this pilot and are provided with suggestions about researchers highly related to the scientific topics their journals and/or issues cover.

2.3.4.4 Requirements

Requirement ID	Description	Priority [core/ essential/ desired][18]	Related Stakeholder Requirement ID
F_PR4.1	The user (researcher or publisher) can request for suggestions of scientific collaborations with researchers.	core	F_Rsr_15
F_PR4.2	The user (researcher or publisher)	core	F_Rsr_15

[18] **Core**: Requirements without which the pilot will not be able to deliver its main objective.
Essential: Requirements for which a short-term work-around could be developed, but over the long run, the requirements have to be there
Desired: Requirements which enrich the offered pilot services but without which the pilot will operate finely.
The pilot must be delivered with all its core requirements and a good portion of essential ones represented, with the overall plan to implement the remaining essential requirements in the following iterations.

Requirement ID	Description	Priority [core/ essential/ desired][18]	Related Stakeholder Requirement ID
	can view the generated suggestions of scientific collaborations with researchers in the form of a list of names.		
F_PR4.3	The suggested collaborations among researchers should be on the basis of the degree of relevance of the research topics and fields they work on, as indicated across their published work and/or their participation in research communities. They should also take into consideration their former and existing collaborations and social-media declared relationships.	core	F_Rsr_16, F_Rsr_17
F_PR4.4	The researcher can request for suggestions of scientific collaborations with research groups and/or research communities.	essential	F_Rsr_15
F_PR4.5	The researcher can view the generated suggestions of scientific collaborations with research groups and/or scientific communities in the form of a list of groups and/or communities highly related to their work.	essential	F_Rsr_15
F_PR4.6	The suggested collaborations between researchers and research groups and/or communities should be on the basis of the degree of relevance of the research topics and fields they deal with, as indicated across the researchers' published work and/or their participation in research communities and the communities' and groups' scientific topics and members profile (if available). They should also take into consideration the researchers' existing collaborations with research groups and communities.	essential	F_Rsr_16, F_Rsr_17
F_PR4.7	Each of the suggested scientific collaborations with other researchers is accompanied by an indicative list of the scientific papers which are highly related to the user's work (if researcher, then the published scientific work and if publisher, then the journal's topics).	essential	F_Rsr_18
F_PR4.8	Each of the suggested scientific collaborations with other researchers is further accompanied by : a. (in case of a researcher initiating the request) the	desired	F_Rsr_18

Requirement ID	Description	Priority [core/ essential/ desired][18]	Related Stakeholder Requirement ID
	scientific topics, fields and/or areas in which they share common interests with the researcher b. (in case of a publisher/editor initiating the request) the scientific topics, fields and/or areas that the researchers work on which overlap with the ones of the publisher's journal/issue		
F_PR4.9	Each of the suggested scientific collaborations with researchers is accompanied by a percentage indicating how strong the suggested collaboration is.	desired	F_Rsr_18
F_PR4.10	Each of the suggested scientific collaborations with research groups and/or communities is accompanied by the scientific topics, fields and/or areas that they cover and are common with the researcher's interest.	desired	F_Rsr_18
F_PR4.11	Each of the suggested scientific collaborations with research groups and/or communities is accompanied by a percentage indicating how strong the suggested collaboration is.	desired	F_Rsr_18
F_PR4.12	The researcher can request for receiving regular notifications about suggestions of scientific collaborations with researchers, research groups and/or communities via e-mail.	desired	F_Rsr_19
F_PR4.13	The researcher can receive and view e-mail notifications about suggestions of scientific collaborations with researchers, research groups and/or communities regularly (e.g., once biweekly).	desired	F_Rsr_19
F_PR4.14	The researcher, who is among top scientist in a specific field (as indicated by pilot 3 and/or pilot 5), is suggested with leading a new, automatically formed research community in the research topic/field s/he excels at.	desired	F_Rsr_20
F_PR4.15	Researchers are informed about new, automatically formed research communities within their research topic(s)/field(s).	desired	F_Rsr_20

Requirement ID	Description	Priority [core/ essential/ desired][18]	Related Stakeholder Requirement ID
F_PR4.16	The profile should be used to automatically suggest possible research collaborations at the international and domestic level.	essential	F_Rsr_2
F_PR4.17	The funder should be able to view a list of experts whose research work is highly related to the one under application for funding in order to select evaluators for the submitted work.	desired	F_Fnd_7

Table 9: Pilot 4 requirements.

2.3.5 Pilot 5: Scientific Field-aware, Productivity and Impact-oriented Enhanced Research Evaluation Services

2.3.5.1 Purpose

To introduce, produce and track new objective metrics of research and scientific performance, beyond conventional metrics associated with conventional indices and impact factors.

Funding agencies, institutions that employ scientists, and scientists themselves, all have a desire, and need, to assess the quality and impact of scientific outputs. It is thus imperative that scientific output is measured accurately and evaluated wisely. Thus, there is a pressing need to improve the ways in which the output of scientific research is evaluated and all the parties involved are encouraging improved practices and methods in research assessment. Such steps are beginning to increase the momentum toward more sophisticated and meaningful approaches to research evaluation.

Outputs other than research articles (such as datasets) will grow in importance in assessing research effectiveness in the future, but the peer-reviewed research paper will remain a central research output that informs research assessment. (San Francisco Declaration on Research Assessment)

Rating the quality and strength of scientific output is an important part of the process of evaluating researchers and institutions, taking decisions about funding allocation and research direction and even decide about the evidence-based health practices and policies. Scientists often arrive at their own judgments about the soundness of research output or technology assessments. Such judgments may differ considerably in the sophistication and lack of bias with which they were made. (Lohr 2004) Tools that meet acceptable scientific standards can facilitate these grading and rating steps and should be based on

- Quality measures: An expanded view holds that quality concerns the extent to which a study's design, conduct, and analysis have minimized biases in selecting subjects and measuring both outcomes and differences in the study groups other than the factors being studied that might influence the results
- Quantity measures
- Consistency measures

The main objective of this pilot is to introduce, produce and track new objective metrics of research and scientific performance, beyond conventional metrics, associated with conventional indices and impact factors. This new type of sciento-metrics enables research sponsors, funding authorities and governmental agencies to shape their research strategies, researchers to be evaluated based on a multitude of factors representative of their productivity, impact and domain rather than through simplified means such as the number of citations within a time period, important research work, in terms of potential, to be brought forward, among others.

2.3.5.2 Key Stakeholders involved

Stakeholder	Description
Researcher	View the evaluation of researchers, papers, data sets, journals, research groups, communities, academic and/or research institutions, countries View their own evaluation both as researchers and for their published research work Follow other researchers, papers, data sets, journals, research groups, institutions in terms of their evaluation View the top researchers, scientific work, publishing means, groups, institutions, countries based on a wide range of evaluation metrics and for specific research topics, fields and/or areas and rank them based on these metrics
Publisher	View the evaluation of their journal and journal issues View the evaluation of researchers, papers, data sets, journals, research

	groups, communities, academic and/or research institutions, countries based on the evaluation metrics they have chosen
Research Agency	View the evaluation of researchers, papers, data sets, journals, research groups, communities, academic and/or research institutions, countries based on the evaluation metrics they have chosen
Research Sponsor and/or Funding Authority	View the evaluation of researchers, papers, data sets, journals, research groups, communities, academic and/or research institutions based on the evaluation metrics they have chosen
Academic and/or Research Institution	View the evaluation of their institution, groups, researchers View the evaluation of researchers, papers, data sets, journals, research groups, communities, academic and/or research institutions, countries based on the evaluation metrics they have chosen

Table 10: Pilot 5 key stakeholders.

2.3.5.3 Overall Description

This pilot focuses on enriching existing scientific research evaluation metrics and introducing new ones in order to overcome the limitations of current evaluation systems and reflect the actual quality, novelty and impact of the published work. By exploiting the full range of openly accessible repositories of scientific work, this pilot aims at establishing an evaluation process which includes aspects such as the evaluation of the source (paper and/or researcher) of a scientific work's citation, the recorded potential interest (going beyond citations and including number of views and downloads among others), the maturity and the crowdedness of the field to which the scientific work belongs, the dynamics both of the research field and the citations of the work as well as the reachability of the work's dissemination means to the rest of the scientific world (e.g., published on authors' websites, abstracting and indexing services, etc). This evaluation process will cover not only published scientific papers, but also data sets, researchers, research groups, communities, academic and/or research institutions and countries.

Within this context, researchers will be able to access their own evaluation – related to them as researchers, their scientific work (papers, data sets, etc), their research group and institution, publishers will be informed about their journals and journal issues evaluation, and members of academic and/or research institutions will be able to monitor their institution's as well as its research group's progress in scientific publishing. Moreover, all users will be able to specify research topic, field and/or domain and view the evaluation of the top k (k may be set as a parameter) published papers, data sets, researchers, research groups, institutions and countries in them. Users will also be able to access the evaluation for a piece of scientific work as well as a scientist, a group, a community, a university, a research institute and/or a country they specify.

Evaluation of scientific output or researcher

There is growing awareness in research communities, government organisations and funding bodies around the concept of using evaluation metrics to analyse research output. A wide range of research metrics are now available to quantify notions of academic impact, profile and scientific quality. However, researchers from a range of disciplines are increasingly questioning the validity and reliability of these analytical tools.

The presence and number of citations are frequently used to assess the influence of a particular article, author, journal or field of research. While it is acknowledged that the number of citations do not necessarily correlate with article quality, nevertheless a high number of citations for a particular article is suggestive of utility by other researchers and as such is one example of a measure of academic impact. However, it does not necessarily provide evidence of ***research impact*** which is characterised by '***an effect on, change or benefit to the economy, society,***

culture, public policy or services, health, the environment or quality of life, beyond academia'. Thus, no available metric of research impact currently exists.

It is important to focus upon the concept of scientific impact. If an article is receiving a small number of citations, it does not mean that it is not being read. There are examples of articles that present a surprising high number of downloads while at the same time they have received only a few citations. However, these articles may be used as reference in practice and therefore it is recommended reading for many undergraduate, postgraduate students or clinicians resulting in consistently high downloads (e,g, articles describing a new surgical technique). In fact, web usage and social networking metrics may ultimately be of benefit especially in the case of articles that are likely to be accessed by health professionals who would rarely offer a traditional citation, yet who seek to apply evidence to their practice. According to the Economic and Social Research Council, "research dissemination does not equal impact". Web usage statistics offer complementary methods in which the academic impact of research may be demonstrated.

A study by an expert in information sciences showed that factors related to higher citation rates included:

1. Articles in the ***English language***.
2. ***Generalist areas*** rather than specific applied disciplines.
3. ***Review articles*** rather than original research.
4. Longer rather than shorter articles.
5. Articles regarding ***established rather than emerging*** disciplines.
6. ISI-indexed journals.

A number of research institutions also acknowledge that lower citation rates may be seen in ***recently published outputs, and between different fields of research***. Different fields of research exhibit quite different citation rates or averages, and the difference can be as much as 10:1. The average 10-year-old paper in molecular biology and genetics may collect 40 citations, whereas the average 10-year-old paper in a computer science journal may garner a relatively modest four citations. Even within the same field, one should not compare absolute citation counts of a ten-year-old paper with those of a two-year-old paper, since the former has had more years to collect citations than the latter. However, these factors are often not taken into consideration when we measure the number of citations. In fact, concerns have been expressed regarding the over-reliance by decision-makers on ***crude citation data*** alone in academic recruitment, promotion and tenure decisions. Moreover, citations take months or years to build up limiting their value as early indicators of impact.

Questions are often raised about ***self-citation and citation circles***, in which a group of researchers agree to cite one another to boost their citation totals. Self-citation is a normal and normative feature of publication, with 25 percent self-citation not uncommon or inordinate in the biomedical literature, according to several studies. It is only natural that when a researcher works on a specific problem for some time, he or she would cite earlier publications from this effort. For example, a researcher in the field of medicinal plant biology, may exhibite 75 percent self-citations in his publications because he might be virtually the only person working in this specific area. Thus, the percentage of self-citations of a researcher should be weighted according to the size of the particular scientific field.

It should be stressed out that the calculation of the number of citations has nothing to do with the understanding of how a research work is being used and perceived. The value of estimating how the reference is being used is indispensable in evaluating the impact of this work. Thus, a paper can be cited in a number of ways as follows:

- ***Passive citation:*** The citing paper may be referring to it amongst several grouped references within a literature review
- ***Positive citation:*** The citing paper may praise the quality of the research or see the article as influential to new understanding

- ***Comparative citation:*** The citing paper may use the research as a benchmark against which to compare their own.
- ***Negative citation:*** The citing paper may argue against the findings of the research or the research methodology

This qualitative review of citations can only identify the esteem in which this research is held, and the true extent of influence on the field.

The most appropriate search engine should be used for searching for articles or identifying citations. The Web of Science, for example, can reach articles as far back as 1900, so is essential when searching for historical content. It can only access journals listed with the Institute for Scientific Information (ISI). Unfortunately this precludes several journals. Other search engines such as Elsevier's Scopus cover a broader range of journals but citing references are only captured for articles published from 1996 and beyond. It is very easy to use, but does require personal or institutional subscription. Google Scholar however, is free to use, also accesses from 1996, and tends to identify more material from non-traditional sources such as government publications. It is vital to note that these different search engines will all identify unique citation material for individual articles, so it is worthwhile using a combination approach.

Another important determinant that has to be taken into account when evaluating a manuscript is the type of publication (Reviews vs original research vs short communications vs letters, basic research vs clinical studies). Each manuscript needs to be evaluated according to its ***type*** and its ***particular scientific field***. Different quality criteria have been proposed according to the type of publication.

The proposed criteria for grading study quality according to the type of clinical study are shown below (Lohr 2004):

- ***Systematic reviews***: study question, search strategy, data extraction, study quality, data synthesis, funding
- ***Randomized Control Trials:*** study population, randomization, blinding, interventions, outcome measure, statistical analysis, funding
- ***Observational trial:*** Comparability of subjects, interventions, outcome measure, statistical analysis, funding
- ***Diagnostic test studies:*** study population, adequate description of test, reference standard, blinded comparison of test, avoidance of verification bias

Scientific research is increasingly a collaborative, multidisciplinary, multi-location and multi-funded activity; upward trends in paper author number are, at least in part, a reflection of this. Thus, an increasingly important problem in research evaluation is how to **"measure" the role of different researchers or different institutions in a research product.** Level and nature of contributions vary and in some cases there is honorary authorship. However, researchers rarely distinguish who is responsible for how much of the work reported. Often no lead author is indicated, and when one is, in some fields the first name listed is traditionally the lead, but in other fields it is the last name listed. Consider a paper by four scientists that has been cited 40 times. If all authors are considered equal contributors then they should receive credit for one-fourth of the paper and 10 citations, instead of receiving a whole publication count and credit for all 40 citations. Fractional counting based on the percentage of contribution of each author to the final research product could be the optimal solution. A well-known criterion for determining the contribution of each author is the ranking of a name in the author's list versus the number of authors. Novel, peer-integrated criteria about the evaluation of the role of each researcher are needed.

An index such as the ***Journal Impact factor*** that has been created mainly to evaluate the performance of scientific journals has been also widely used to evaluate individual researchers. Major concerns have been expressed regarding the misuse of the impact factor to assess an individual researcher or individual publication quality, rather than considering individual citation rates. In fact, evaluating authors in terms of their JIFs has been equated as to evaluating university student applicants in terms of the average marks of the secondary schools from which

the applicants have graduated, instead of comparing them in terms of their own individual marks.

Indeed some authors suggest that impact factors and citation data are notoriously unreliable, and argue that the actual citation rate of a journal does not reflect the value of an individual paper. On the other hand, analysis of citations can provide an important indication of the academic impact of an article, including ***information regarding the citing authors, their institutions and the countries of origin***. This analysis is able to suggest how local, regional or global the reach of the research actually is.

A relatively simple measure of an individual researcher's impact in terms of citations is the h-index. The h-index calculates the highest number of articles published by the author that have the equivalent number of citations or above. Thus, it is an index that attempts to measure ***both the productivity and impact of te published work*** of a scientist or scholar. For example, an h-index of 10, would suggest an experienced researcher with 10 articles that have at least 10 citations each. There are a number of situations in which *h-index* may provide misleading information about a scientist's output:

- The *h*-index discards the information contained in author placement in the authors' list, which in some scientific fields is significant
- The *h*-index is bounded by the total number of publications. This means that scientists with a short career are at an inherent disadvantage, regardless of the importance of their discoveries. For example if Albert Einstein died after publishing his four groundbreaking Annus Mirabilis papers in 1905, his *h*-index would be stuck at 4 or 5.
- The *h*-index does not account for the typical number of citations in different fields. Different fields, or journals, traditionally use different numbers of citations.
- The longevity of the h-index is a major weakness, as it can never reduce even if a researcher retires or dies. Thus, it is a reliable tool to evaluate researchers only in the same stage of their careers.

Various proposals to modify the *h*-index in order to emphasize different features have been made (*m*-index, g-index, c-index). As the variants have proliferated, comparative studies have become possible and they demonstrate that most proposals do not differ significantly from the original *h*-index as they remain highly correlated with it.

Many of the citation analysis tools were developed with a traditional pre-digital era in mind. A major problem of all these traditional, pre-digital scientometric measures is that they are individual one-dimensional metrics, while evaluation of research output and impact is a multi-dimensional process. The task of capturing scientific impact in our modern digital era requires new impact measures to be created, based on ***social network analysis and web usage log data***. Furthermore, the online publishing medium and the advantages inherent in Open Access content provide tremendous potential for the development of alternative measures of research impact and influence. In fact, a few open-access journals, such as the PLOS One, are already experimenting by providing information relating to online usage, downloads, citation activity, blog and media coverage, commenting activity and social bookmarking, aiming to investigate how these indices could create new standards to measure the "impact" of research.

An example of these new scientometric measures is shown in Figure 8.

Figure 8: Example of scientometric measures currently in use.

Evaluation of Journals

The Journal Impact Factor (***JIF)*** is now commonly used to measure the impact of journals and by extension the impact of the articles they have published. Journal impact factors are calculated annually, defined as the average number of citations in ISI-indexed journals to those papers that were published in the preceding two years. There has been widespread criticism of the impact factor rating system, centred around validity, manipulation and misuse. ***Validity concerns*** relate to the fact that the impact factor cannot be reproduced independently, and may be heavily influenced by author and journal self-citation rates.

Researchers attribute the relatively low impact factor ratings of some journals, to the ***restrictive two-year timeframe*** for collection of citations in relation to the scientific field of interes . In cutting edge, fast-moving fields such as molecular biology, for example, one would expect many citations within the first two years after publication. However, in other fields such as radiology, the impact of an article is often more of a 'slow burn', gaining significant numbers of citations often several years after publication. Editorial policies can also influence citation rates, for example by increasing the percentage of review articles accepted, or encouraging journal self-citation.

The skewed distribution of citations across all papers in a journal means that the impact factor is not a reflection of the average citation rate of any given paper within the journal. Few papers in a journal are therefore cited at approximately the impact factor rate. Indeed, even in a high impact factor journal, a small number of papers carry huge influence. In 2005, 89% of Nature's impact factor was generated by only 25% of the articles.

For this reason, alternative journal ranking systems have been proposed which include a weighting factor on their citations. An example is the SJR (SCImago Journal Rank), an independently-derived metric using Scopus data, which has far greater coverage than Web of Science and can be calculated for journals which are not currently ISI-listed.

The SJR considers the average number of weighted citations received in the selected year by documents published in the previous three years (as opposed to two years for ISI impact factor calculations). However caution should be taken in comparing journals using the SJR, as a number of errors of categorisation were noted during a recent search. For example, Radiography was listed in the 'Pathology and Forensics' category, while Clinical Radiology was listed as an 'Oncology' journal, rather than in the 'Radiology' category.

A number of different metrics are offered on the SCImago Journal and Country Rank website and also using the Scopus Journal Analyzer tool:

1. ***SJR***, as described above
2. ***SNIP***, Source normalized impact per paper. Corrects for differences in frequency of citations across different research fields
3. ***Citations***, total citations received by journal in the year
4. ***Docs***, total number of documents published by journal in the year
5. ***Percent not cited***, Percentage of documents published in the year that have never been cited to date
6. ***Percent reviews***, Percentage of documents in any year that are review articles

An example of evaluation of four journals from different scientific fields is given in Figure 9. The journals examined were Pharmacology and Therapeutics (P&T), Basic Research in Cardiology (BRC), European Journal of Endocrinology (EJE) and European Journal of Cardio-thoracic Surgery (EJCTS). According to an evaluation based on the the SJR metric, it is clear that P&T had a higher ranking than BRC, which has a higher ranking than EJE and EJCTS. Almost identical results come up if instead of SJR, we used the JIF. However, if the evaluation is based on the SNIP index, which corrects for differences in frequency of citations across different research fields, P&T has the highest ranking while BRC, EJE and EJCTS have comparable rankings.

Figure 9: Example of evaluation of four journals.

Since scientific literature is now mostly published and accessed online, a number of initiatives have attempted to measure scientific impact from usage log data. The resulting usage data allows scientific activity to be observed immediately upon publication, rather than to wait for citations to emerge in the published literature and to be included in citation databases such as the JCR; a process that with average publication delays can easily take several years. In fact, a ***Usage Impact Factor*** which consists of average usage rates for the articles published in a journal, similar to the citation-based JIF has been proposed. However, this index does not overcome the shortcomings of JIF, such as validity concerns, restrictive timeframe, skewed distribution of usage citations.

Towards novel criteria, parameters and indicators of research assessment

Various new metrics will be provided by the OSL Platform to all registered users.

Citation Activity and Evaluation in the Scholarly Literature: This is a well established metric, and one that most people agree can be viewed as an indicator of the worth of an article. A new metric, which will give a qualitative flair to the strictly quantitative evaluation of research, is through analysis of the citations the research work has. The tool will permit the number of citations to be weighted against the value of these citations. ***Citation analysis*** can provide important information, including ***evaluation of the citing authors, their institutions and the countries of origin***. This analysis is able to suggest how local, regional or global the reach of the research actually is. Additionally, ***the timeliness and density of the citations*** can be used as an evaluation metric.

Another important factor is connected with ***the dynamics of the field*** that the research work resides in, i.e. whether the research field is in its infancies, at its peak or in decline. According to this metric, ***the impact of the research work is weighted depending on the research potential and dynamics of the specific field***.

The measured impact of the research work is also affected by the level of difficulty for other researchers to have access to this work. Self-archiving authorization, publisher archiving provisions, copyright provisions, abstracting and indexing services, and reference linking, are some of the factors that affect this metric. Moreover, citations are determined by the scientific area of interest. For example, generalist areas tend to have increased number of citations compared to specific applied disciplines. Thus, the platform will be able to calculate the ***total number of citations per paper compared with the average citations per paper in the same field over the same time period. This index should also be weighted against the type of publication (review vs original research vs letter vs brief communication).***

A decisive factor of citation evaluation will be also based on the research trends analysis which will be part of the platform. Thus, it is known that articles on ***established rather than emerging disciplines*** are getting more citations on average. The platform will be able to calculate this effect and weight the number of citations against it.

To examine the relative citation score of an individual researcher, ***the citation score for each publication*** should be calculated and compared to ***expected citation score*** based on the scientific field, time and type of publication, the contribution of the researcher etc. The next step is to make a ratio of the two to gauge ***the percentage of publications that belong to better than average, average, or lower than average***. This index is important in order to weight the productivity of a researcher (as defined by the number of publications per year) against the quality of this work. Especially in recent years, it is increasingly accepted that productivity and quality of research usually are one against the other. In fact, production of high numbers of publications per year is often linked to poor quality and/or lack of innovation.

Citation-agnostic metrics: In this category fall a number of metrics related to the quality and not the quantity of the scientific work independent of the citation analysis. These quality measures pertain the ***scientific focus of a researcher, the publication of reviews in relation to its main scientific focus and the innovation and potential introduced by a scientific work***. In fact, if the publications of a researcher cover a very wide area of scientific fields or are too much focused, this is a low-quality characteristic. OSL platform by semantically mapping the publications of a researcher will be able to construct a graph representing the degree of research focus. Furthermore, the platform will be able to search and find published reviews of the researcher in relation to its main scientific focus. These publications are considered as an indication of significant contribution and prestige for the researcher. Finally, innovation is a difficult quality to automatically spot and evaluate in a publication. The approach proposed by OSL platform could include the calculation of 2 different indices; an ***index of novelty*** and an ***index of novelty and potential interest***. Both indices could be calculated for a research work or for a researcher. The steps for the calculation of these indices are described below:

- Define 2 semantically-linked scientific fields related to the evaluated publication or researcher e.g. "thyroid hormone" (field A) and "myocardial infarction" (field B)
- For calculation of the ***index of novelty,*** define the ranking of field A in the field B and the ranking of field B in the field A according to the number of publications.
  e.g. the ranking of "myocardial infarction" in the field of "Thyroid hormone" is shown below

"Thyroid hormone" (Field A)

1. Thyroiditis
2. Metabolism
3. Hyperthyroidism
4. Nuclear receptors
5. Transcription factors
6. Liver
7. Pregnancy
8. Iodine
9. Thyreoidectomy
10. Antidodies
11. Body weight
12. Neoplasm
13. Hepatitis
14. Signal transduction
15. Muscles
16. Thyroid neoplasms
17. RAR
18. TRH
19. Thyroglobulin
20. Pituitary gland
21. Brain
22. Anti-thyroid agents
23. Growth hormone
24. Graves disease
25. Homeostasis
26. Kidney
27. Insulin

28. Lipids
29. Cell differentiation
30. Myocardium
31. Steroid receptors
32. Estrogen
33. Glucose
34. Glucocorticoids
35. RXR
36. Surgery
37. Autoimmunity
38. Cholesterol
39. Cell proliferation
40. Congenital hypothyroidism
41. Adipose tissue
42. Skeletal muscle
43. Vitamins
44. Aging
45. MAPK activity
46. Hypothylamus
47. Thyroid hormone resistance syndrome
48. Prolactin
49. Xenopus
50. Metamorphosis
51. Intestine
52. Somatotropin
53. Immunity
54. Apoptosis
55. Lung
56. Brain development
57. Depression
58. Estrogen receptors
59. Pressure
60. Testosterone
61. Lactation
62. Mitochondrion
63. Arteries
64. Growth and development
65. Lipoproteins
66. Peroxisomes
67. Cytokines
68. Myosins
69. Eythyroid sick syndrome
70. Erythrocytes
71. Stem cells
72. Vitamin D
73. Epithelial cells
74. Hypocambus

75. Epithelium
76. Heart failure
77. Hashimoto disease
78. Adipose tissue
79. Breast neoplasms
80. Amiodarone
81. PPARs
82. Thermogenesis
83. Hepatocellular carcinoma
84. Mutagenesis
85. Gonads
86. Leptin
87. Exercise
88. Diabetes
89. Dialysis
90. EGF
91. Anatomy
92. TNFa
93. Neuroglia
94. Insulin resistance
95. Learming
96. Androgen receptors
97. PTH
98. GRH
99. Postmenopause
100. Morphogenesis
101. Osteocalcin
102. AP-1
103. Atherosclerosis
104. PKC
105. Atrial fibrillation
106. Catalase
107. Myocardial Infarction
108. Kidney failure
109. Skin
110. Prostate

"myocardial infarction" ranking in the field of"Thyroid hormone" ***107***

"Thyroid hormone" ranking in the field of "myocardial infarction" ***215***

Index of novelty= 107+215 / 2= **161,** **$\log_2$=7.3**

- For calculation of the ***index of novelty and potential,*** define the total number of publications (NoP) in field A and in the field B separately and the total number of publications in the combined field A+B.

e.g. In the field of "thyroid hormone" ***NoP=21,566***

In the field of "myocardial infarction" ***NoP=100,000***
In the combined field of "thyroid hormone" and "myocardial infarction"
NoP=112

Index of novelty and potential interest= 21,566+100,000 / 2*112= 542, $\boldsymbol{log_{10}=2.73}$

PATENTS: Generation of Patents is also defined as an additional research product linked to the quality, novelty, usefulness and impact of research to society. OSL platform will provide the tools to search inside the world's important patent databases and indentify patents linked to the work of a researcher. OSL aims to build advanced text mining algorithms to bring out insights in minimum time which would take days for a researcher. Patent analysis is needed to estimate the quality of a patent based on the collection of field-specific patents and the extraction of relevant quality performance indicators. Patent analysis tools will be able to:

- Represent technology trends over time related to a scientific field
- Generate standard patent reports and maps within a scientific field
- Identify key inventors within a particular scientific field
- Identify new application areas, new licensing and research opportunities
- Identify patent/company acquisition opportunities, and formulate potential collaborations between academia and industry.

Publication usage and scoring: The work will be evaluated based on the potential interest about it, the number of other researchers who downloaded it and their own research evaluation. The platform will allow registered users to leave notes directly on parts of the work, comments on the entire article, and to make "ratings" about content itself. ***Ratings*** could be categorized as ***outstanding, positive, neutral or negative work***. According to these ratings, graphs could be constructed showing how a scientific paper is perceived by the community. In fact numerical ratings are used by multiple services in the outside world to gather user feedback on the worth of a product or service. Therefore, it seems obvious that allowing users the ability to leave this type of feedback would be another useful metric when evaluating articles. The ***organization of surveys among groups of experts*** to evaluate research work based on specific quality criteria (e.g. study population, randomization, blinding, interventions, outcome measure, statistical analysis, funding for RCTs) could be possibly implemented with the OSL platform.

Comments and Notes: This functionality will be build-in the OSL platform, and represents a way to have conversations about articles; to gain an indication of what other users feel about those articles. Therefore, evaluation of the commenting and notation activity of any article is an indication of the interest it has generated in the community.

Social Bookmarking and Blog Coverage: It is reasonable to consider that an analogy exists between citing an article (when someone values an article enough to cite it in their own publications) and bookmarking an article (when someone values an article enough to save it in a publicly viewable location for future reference). OSL platform will provide a count of how many people have bookmarked the paper at each service. An example of such service is CiteULike. In addition, Blog coverage is closely related to media coverage and indicates how "newsworthy" or "interesting" an article might be for a readership that is wider than normal.

Based on the combination of the above metrics, the Platform will be able to calculate the overall evaluation of a research work.

2.3.5.4 Requirements

Requirement ID	Description	Priority [core/ essential/ desired]	Related Stakeholder ID (if any)
F_PR5.1	The user can request for viewing the evaluation for a specific paper, data set, journal and/or researcher s/he chooses.	core	F_Fnd_2, F_Fnd_3, F_Rsr_12
F_PR5.2	The user can request for viewing the evaluation for a specific research group and/or community s/he chooses.	essential	F_Fnd_4, F_Rsr_12
F_PR5.3	The user can request for viewing the evaluation for a specific academic and/or research institution and country s/he chooses.	desired	F_Fnd_5, F_Rsr_12
F_PR5.4	The user can request for viewing the evaluation for a specific academic and/or research institution and country s/he chooses for a specific domain.	desired	F_Fnd_5, F_Rsr_12
F_PR5.5	The user is presented with all the available evaluation metrics applied in each case (paper, data set, journal, etc).	essential	F_Rsr_12.1
F_PR5.6	The user is able to choose among the available evaluation metrics based on which s/he prefers to view evaluation results in each case (paper, data set, journal, etc).	desired	F_Fnd_6, F_Rsr_12.1, F_Prs_3, F_Prs_4
F_PR5.7	The user is able to choose a specific research topic, field or area and view the evaluation of its top subjects and objects (i.e., researchers, papers, data sets, etc) based on the available metrics.	desired	F_Rsr_4, F_Prs_3, F_Prs_4
F_PR5.8	The user is able to rank the evaluation subjects and objects (i.e., researchers, papers, data sets, etc) in a specific research topic, field or area based on each evaluation metric.	desired	F_Rsr_12.2, F_Prs_3, F_Prs_4
F_PR5.9	The researcher can access, at any time, their own evaluation along with the current evaluation of their published papers and data sets as well as of their research group and/or institution.	essential	F_Rsr_21
F_PR5.10	The publisher can access, at any time, the evaluation of their own journals, journal issues, published papers and data sets.	desired	F_Pbl_28
F_PR5.11	The user may choose to follow the evaluation of specific papers, data sets, journals, researchers, research	desired	F_Rsr_22

	groups, communities and institutions.		
F_PR5.13	The evaluation criteria and methodology should be publicly available and should be ideally based on measurements that may be drawn by accessing publicly accessible data sources, or publicly accessible statistical information and reports.	Core	F_Rsr_13

Table 11: Pilot 5 requirements.

2.4 Legal analysis of the user requirements

2.4.1 Introduction

The following chapter provides an initial legal analysis of the user requirements presented in the previous sections of this deliverable. A more specific description of the technology that will be used to achieve the desired goal will be necessary to properly assess the legal impact of the requirements. Such detailed legal analysis will be prepared in Deliverable 3.2: "Legal and IPR Management Framework Specification" and a full legal evaluation of the project developments and results will be performed in WP8.

The most important legal aspects that have to be taken into account during the development of the OpenScienceLink project pertain to (1) the privacy and data protection regulations, as well as (2) the intellectual property framework.

Based on the initial draft requirements as presented, this deliverable will focus on the legal concepts that will inevitably have an impact on the requirements, regardless how these will change during the project. These concepts will be presented shortly. For more information, the reader can consult Deliverable 3.2.

The most important legal aspects that have to be taken into account during the development of the OpenScienceLink project pertain to (1) the privacy and data protection regulations as well as (2) the intellectual property rights framework.

2.4.2 Privacy and protection of personal data

2.4.2.1 Introduction

In the course of the development of the OpenScienceLink platform, it is likely that a number of privacy issues will be encountered. In this chapter, we will explain (1) the applicable legal framework, (2) what constitutes a personal-data processing operation, (3) the concept of 'data controller' and (4) how to apply these concepts in the OpenScienceLink project. At this stage, we can identify two possible data-processing operations: (1) the processing of informationfrom the user profile and (2) the processing of clinical and biomedical data.

2.4.2.2 Legal framework

The most important privacy and data protection regulation in the European Union is laid down in the following documents:

- Article 7 (Respect for private and family life) and Article 8 (Protection of personal data) of the Charter of Fundamental Rights of the European Union[19];

[19]Charter of the Fundamental Rights of the European Union (2000/C 364/01), http://www.europarl.europa.eu/charter/pdf/text_en.pdf

- Article 8 of the European Convention of Human Rights (Right to respect for private and family life)[20] and;
- Directive 1995/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data[21] (Data Protection Directive).

These three instruments define the rules and principles which apply to the processing of personal data in the European Union. For the purpose of this deliverable, the focus will be put on the Data Protection Directive, as this directive contains concrete requirements with which the prototypes have to comply.

Note that there are comprehensive reform proposals launched for a new Data Protection Regulation [22] in the beginning of 2012. This Regulation is currently not yet final. However, the proposed changes relevant for OpenScienceLink will be followed up throughout the OpenScienceLink project.

2.4.2.3 Processing personal data

One of the main questions that have to be answered is whether the development and implementation of the OpenScienceLink prototypes would lead to 'processing of personal data' of users, content providers, patients, etc.

Article 2 of the Data Protection Directive delineates the precise scope of the concept of personal data. "Personal data" constitutes:

'…any information relating to an identified or identifiable natural person ('**datasubject**'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to this physical, physiological, mental, economic, cultural or social identity.'

It is evident that the definition of 'personal data' is quite broad, which makes it hard to determine when exactly we are dealing with personal data. Moreover, the scope of 'personal data' is subject to constant transformation due to societal, cultural, technology or economic changes. The interpretation of 'personal data' also depends on the local practice of the Courts of the different Member States.

For the Data Protection Directive to apply there must be some form of 'processing' of the personal data. This includes collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction.[23] By including storage in the definition, the mere acts of 'holding' personal data or 'accessing' the information constitute a processing operation subject to the rules of the directive.

OpenScienceLink

The scope of the Data Protection Directive is very broad and it will most likely apply to the collection, storage and use of data in the OpenScienceLink projects, e.g. in the case of the creation of user profiles.

[20] European Convention of Human Rights, http://www.echr.coe.int/Documents/Convention_ENG.pdf

[21] Directive 95/46/EC of the European Parliament and of the Council of 24.10.1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive), OJ L 281, 23.11.1995.

[22] European Commission, Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Brussels, 25 January 2012, COM (2012), 11 final.

[23] Data Protection Directive, article 2 (b).

2.4.2.4 Data Controller

In the OpenScienceLink environment it is likely that the tools developed based on the user requirements, would allow access to personal information. Consequently, the Data Protection Directive will apply and in this context, it is crucial to first define the 'data controller' to know who will be responsible for the application of the Data Protection Directive.

The *controller* is, according to Article 2 (d) of the Data Protection Directive, the natural or legal person, public authority, agency or any other body which alone, or jointly with others, determines the purpose and means of the processing of personal data. It is important to identify who the controller of any processing is, since the controller is the one liable for the legality of the processing and the fulfilment of the obligations towards the national data protection authority and the data subjects.

The concept of data controller and its interaction with the concept of data processor play a crucial role in the application of the Data Protection Directive, since they determine who shall be responsible for compliance with data protection rules, how data subjects can exercise their rights, which is the applicable national law and how effective Data Protection Authorities can operate.[24]

OpenScienceLink

Crucial for the development and functioning of the OpenScienceLink platform will be to define who the 'data controller' is. Determining this natural or legal person is crucial,for he or she will be liable for the legality of the processing and the fulfilment of all obligations that come with the applicability of the Data Protection Directive.

2.4.2.5 Legitimate basis for processing personal data

Article 7 of the Data Protection Directive contains an exhaustive list of six legitimate grounds for personal data to be processed:

- The data subject unambiguously given his consent; or
- Processing is necessary of the performance of a contract to which the data subject is party; or
- Processing is necessary for compliance with a legal obligation to which the controller is subject; or
- Processing is necessary in order to protect the vital interests of the data subject; or
- Processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed; or
- Processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such rights are overridden by the interests for fundamental rights and freedoms of the data subject.

The proposed Regulation restricts the processing of personal data for research purposes in so far that such personal data may only be used if the research objectives cannot be reached by using anonymous data and, in that case, the data enabling the attribution of the information to a particular person is kept separately from the other information, to the extent possible (art. 83 Reform proposal). Specific rules are suggested for publication of research results as well.

[24] Article 29 Working Party, Opinion on the concepts of 'controller' and 'processer', (31) 35

OpenScienceLink

For the OpenScienceLink platform, the relevant and most important grounds to process personal data upon will be (1) the consent of the user and (2) legitimate interests.

2.4.3 Data processing operations in the OpenScienceLink project

At this stage, we can identify 2 data processing operations which will potentially occur in the development of the OpenScienceLink platform. If processing of personal data occurs, the obligations from the Data Protection Directive will have to be adhered to.

2.4.3.1 User profile

It is proposed that users of the OpenScienceLink platform would register to the platform by creating a *user profile.* This profile would contain personal data including but not limited to: name, title, position, address, e-mail, main areas of scientific interest and areas of methodological expertise. This data would be used for the purposes of OpenScienceLink, for instance, finding competent and relevant peer-reviewers.

Using this type of information would be considered as accessing and processing personal data. Article 7 of the Data Protection Directive (*supra*) provides us with 7 legitimate grounds on which such data processing operation can occur.

Moreover, the obligations from the Data Protection legislation will have to be adhered to. Such an obligationconstitutes for instance the notification to the Data Protection Authority. Another obligation is the application of the 'purpose specification/limitation' principle. Personal data must be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes.[25]

This condition can be translated into two requirements. First, the purpose must be 'explicit', thus the user has to be informed specifically about the purpose of data collection. Second, once the data is collected, it cannot be used for another purpose that is incompatible with the originally communicated purpose. For the OpenScienceLink end-user, this means that any personal data collected and processed cannot be used for any other purpose than the one that the user had in mind when he provided the data or a purpose compatible with this original one. In most of the cases the original purpose of providing the data would be for example to enjoy full access from the possibilities and functions of the OpenScienceLink platform. A use of the collected data incompatible with the latter one would constitute a secondary use for which a new notification and consent of the data subject would be required. As a consequence, it should be noted that any use of personal data performed without providing the necessary information to the data subject, as described in Article 10 of the Directive, could be considered a breach of this particular provision.

OpenScienceLink

Using the data provided by the data subject in his/her 'user profile' will constitute a form of 'processing personal data'. Consequently, the Data Protection Directive will apply. Data subjects should be clearly informed about the purposes of the possible data collection, and means of processing and the purpose of this data processing operation must be specific and legitimate.

2.4.3.2 Anonymised research data

An especially crucial issue for the use and development of the OpenScienceLink platform is the processing of the clinical and biomedical data as foreseen in the pilot services.

[25] Article 6 (1) b Data Protection Directive

For the purpose of some scientific or scholarly studies, it might be necessary for the data providers to use data about living persons (rather than test animals for instance). To avoid the application of the Data Protection Directive, this data should be anonymised by these data providers. It is crucial that the data remains fully anonymous on the OpenScienceLink platform, otherwise the Data Protection regulations would become applicable, and we would have to comply with its obligations.

In its opinion on personal data, the Article 29 Working Party defines anonymous data as "any information relating to a natural person where the person cannot be identified whether by the data controller or by any other person, taking account of all the means that could reasonably be used either by the controller or by any other person to identify that individual."[26] According to Recital 26 of the Data Protection Directive, "the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable."

Consequently, the processing of anonymous data does not fall within the scope of the Directive and does not have to comply with the legal data protection requirements. However, the process of rendering the personal data anonymous constitutes 'processing personal data' and falls within the ambit of the EU Data Protection Directive.

It is important in the context of the OpenScienceLink project to note that anonymised data in practice might not per se mean that the data subjects can no longer be identified. Especially when data from multiple sources is combined, this may in fact produce data that can be used to trace the identity of the person concerned. This would mean that anonymised data could become indirectly identifying data. A familiar example is an uncommon occupation which when combined with someone's age or the place he lives could make clear who the person concerned is. Note that the re-identification possibilities also depend on who will be able to access the platform and who will be data controllers.[27]

If it is possible to connect the research data with an individual by means of the tools available to a researcher, then the data concerned will amount to indirectly identifying data. This may entail that processing the data will have to be done in accordance with the provisions of the Data protection Directive.

OpenScienceLink

It must be ensured that the clinical and biomedical data provided is fully anonymous, including removing any risk of possible indirect identification of the data subjects. Otherwise, the Data Protection Directive will apply, including all principles and obligations that come with it.

2.4.3.3 Conclusion

At this stage, it is advisable to design the platform in the form of a voluntary service where users would join freely and/or upon invitation and where they would be clearly informed about the purposes of the possible data collection, and means of processing. That way they could express their consent which would greatly limit any possible danger of infringing any regulation in this matter. A strong reduction of the legal risk involved, will ultimately guarantee a successful implementation of the platform in practice.

2.4.4 Intellectual property rights

Another crucial concern during the development and use of the OpenScienceLink platform refers to the intellectual property rights on the content included in the platform. This chapter will focus

[26] Article 29 Working Party, Opinion on concept of personal data, (21) 26

[27] Surfdirect (2009). The legal status of raw data: a guide for research practice. *CIER,* 41 and C. Gideon Emcee, Building a sustainable framework for open access through information and communication technolgoies, IDRC-CRDI, 14.

on 2 important issues which could bring about intellectual property rights concerns; (1) the legal status of the clinical and biomedical research data and (2) the papers for peer-review.

2.4.4.1 Research data

When we talk about raw research data, we talk about bare 'facts' and ideas. This is a subject matter which is generally not protected by any intellectual property rights. However, with the information presented at this stage, the collection of clinical and biomedical research data can possibly fall under two levels of protection.

(1) Intellectual property rights can attach to the original expression of the facts in a particular form in the form of *copyright*.

(2) Collections of scientific data can also be protectable under the European *sui generis database* right if the maker of the database showed a substantive investment.[28]

Copyright regime

General Framework

Attempts to harmonise copyright law can be dated back to the signature of the Berne Convention for the Protection of Literary and Artistic Works on 9 September 1886 (*Berne Convention*). Currently, the norms of copyright are embodied in an interlocking network of (1) international treaties, (2) European regulation and (3) national legislation. Although there are some universal copyright principles, and bilateral, regional and multilateral treaties did push the harmonization of the national rules on copyright, the contemporary copyright norms are still not universal. This makes matters quite complex.

The most important copyright regulations on an international level are laid down in the following documents:

- Berne Convention for the Protection of Literary and Artistic Works (1886, latest version, Paris 1971)[29]
- Rome Convention for the protection of Performers, producers of phonograms and broadcasting organisations (1961)
- Agreement on Trade related aspects of intellectual property rights (TRIPS) (1994)
- WIPO Copyright Treaty (1996)
- WIPO Performances and phonograms Treaty (1996)

The most important copyright regulations in the European Union are constituted by several directives, implemented by the Member States, combined with the decisions of the European Court of Justice. Some relevant examples include:

- Directive 93/98 harmonising the term of protection of protection of copyright and related rights (1993) OJ 290/09 (codified by Directive 2006/116/EC)
- Directive 96/9 on the legal protection of databases (1996) OJ 77/20.
- Directive 2001/29 on the harmonisation of certain aspects of copyright and related rights in the information society (2001) OJ 167/10

The preeminent international treaty on intellectual property rights is the Berne Convention, which dates back to 1886. Countries bound by the Berne Convention are obligated to protect "collections of literary and artistic works such as encyclopedias and anthologies which, by reason of the selection or arrangement of their contents, constitute intellectual

[28] Article 7 of the Database Directive

[29] WIPO, Berne Convention for the Protection of Literary and Artistic Works,at http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html#P85_10661

creations."[30]Consequently, to fall under the protection of the Convention, there are two requirements: (1) there needs to be a minimum form of originality, and (2) this creation has to be expressed in a certain form.[31]

This means that in practice, the Berne Convention may protect certain creative databases, but presumably does not extend to protection for noncreative databases. An "intellectual creation" is required.

Problematic for OpenScienceLink might be the fact that the Convention refers to 'collections of literary and artistic works' which calls into question the protection of databases that consist of non-copyrightable data elements, i.e. research data for instance.

The Convention is moreover not clear about the criterion for protection, how is an "intellectual creation" defined? The Berne Convention does not explicitly mention 'originality' as a criterion. In what form is this expressed? Generally, it is determined that originality or creativity is required for works protected by the Convention. Defining its scope is a matter for the different national Courts in the Member States.[32]

OpenScienceLink
It is questionable whether the collection of non-copyrightable research data will fall under the protection of the Berne Convention, as this Convention protects 'collections of artistic & literary works'.

European Framework

First of all, it should be clarified that the matter of copyright law in Europe is not fully harmonised. This means that there are many differences between countries in the European Union in how they regulate these issues. At the European level, there are however three categories of works which enjoy a harmonized regime; i.e. software, photographs and databases.

In the OpenScienceLink project, we will especially focus on the matter of databases. Raw research data as such cannot enjoy copyright protection. Research data consists of facts, facts on which researchers or scholars can never have an intellectual monopoly. What might possibly be protected by the copyright regime, are the form in which these bare facts are presented.

The applicable criterion for copyright to apply is the criterion of 'originality'. According to Article 1 of the Database Directive, for a work to receive legal protection, it must be 'original', i.e. the author's 'own intellectual creation' by reason of the selection or arrangement of the contents.[33] Whether collections of scientific research data will meet the criterion of 'originality' is a question which will be dealt with on a case-by-case basis. It depends on the interpretation of each national Court. Essential for copyright to apply is that the database shows a certain character or a certain creative or intellectual effort by the author.

In case the structuring, selection or wording of the data would be protected by copyright, it is necessary to define the 'copyright holder'. The copyright holder enjoys certain *exclusive* rights to carry out certain actions, e.g. duplication of publication. The copyright holder is the sole party that can perform these acts; others must secure the consent of this right holder. Please also note

[30] Article 2 (5) of the Berne Convention

[31] K. Janssen & J. Dumortier, The protection of maps and spatial databases in Europe an dht US by copyright and the sui generis right, The John Marshall Journal of Computer and information law, 2006, 2, 199

[32] Ibid.

[33] Article 1 of the Database Directive and *Infopaq International A/S v Danske Dagblades Forening,* C-5/08, 16 July 2009; *Bezpecnostni softwarova asociace v. Ministerstvo kultury,* C-393/09, 22 December 2010; *Eva maria Painer v. Standard Verlag GmbH,* C-145/10, 1 December 2011 and *Football Dataco v. Yahoo UK Ltd.,* C-604/10, 1st March 2012

that the copyright holder is not *per se* the author of the 'work'. Contractual agreements or the involvement of superiors when processing the data and licenses will be relevant here.

In the OpenScienceLink context, it is likely that the arrangement of the research data will be based on standard arrangements needed for interoperability and the only value lies in the fact that the database would be as complete as possible. This leaves little room for originality or an intellectual creation in the selection or arrangement of the data. As a result, scientific databases will in most cases not likely meet the threshold for copyright protection. It would thus be interesting to also look at the *sui generis* right protection, which protects the investment in a database, rather than the creativity.

OpenScienceLink Bare facts are never eligible for copyright protection, but the format in which they are presented could be. Copyright protection requires originality. It is questionable whether a database of research data will be copyright protected. However, this question depends on the way in which the database will be developed. Will there be a certain level of originality in the selection and arrangement of the collected research data?

Sui generis database right

A collection of scientific data may also be subject to protection by the sui generis database right. This is an entirely different matter than copyright. Where the database right is concerned, it is not the creativity or originality of the database or its contents that matter, but the protection of the investment made in order to assemble the collection of data. The database right is specifically intended for someone who invested in the database to recoup his investment by exploiting the database. The database right protects the investor against third parties that wish to copy large portions of the database or to use it without consent or without paying. Database right also applies to non-commercial databases, as long as the requirements explained below have been complied with.[34]

According to Article 7 of the Database Directive, the maker of a database showing a substantial investment (assessed quantitatively and qualitatively) in either the obtaining, verification or presentation of its contents has the exclusive right to prevent the extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database. Like copyright protection, the *sui generis* database right arises automatically, without any formal requirement, at the moment the database is completed or disclosed to the public.

The first criterion to assess in the context of the OpenScienceLink project is the matter of 'substantial investment'. The Court of Justice has given a particularly narrow interpretation to the Directive's requirement that the database show a substantial investment in either the obtaining, verification or presentation of its contents. In a number of landmark cases it was decided that 'obtaining' excludes the costs incurred in the creation of new data (such as generating fixtures lists) from being considered relevant to satisfy this requirement.[35] What can be taken into account are the costs necessary for the verification of the accuracy of the data and for the presentation of such data to third party users.

It is unclear at this stage whether the collection of research data will qualify as a substantial investment under the Database Directive. It is decided on a case-by-case basis and moreover, it depends on how much time, money and effort will have gone into collecting and verifying the

[34]Surfdirect (2009). The legal status of raw data: a guide for research practice. *CIER.* 52.
[35] C-388/02, (2004) Fixtures Marketing, *ECR I*-10497 and C-203/02, (2004) British Horseracing board Ltd, *ECR I*-10415.

data, with that investment being solely in creating the database. Moreover, this threshold will be examined on a case-by-case basis by the national courts.

The second issue to assess is the 'database right holder'. The producer is normally the right holder in respect of the database, he is the party that can grant consent to retrieve and reuse substantial portions of the database or to repeatedly and systematically retrieve non-substantial portions.

In the context of the Open Science Link project, there are several possibilities. The producer of the database is the party that bears the risk in respect of the investment in the database. It is likely that several databases with research data will be combined into the OpenScienceLink platform. Each of these databases, which have required a substantial investment, and thus will each have a database right holder, whose permission is required to use the database. The new database that will be created may or may not constitute a new database allowing for a separate database right.

Practically, if the Open Science Link database would qualify for a *sui generis* database protection, following actions would require consent from the maker of the database:

- copying or downloading substantial parts of the database;
- repeatedly and systematically retrieving non-substantial parts;
- publishing substantial parts of the database.

The following actions would not require the makers' consent:

- using a database for scientific/scholarly research if no substantial parts of the database are published (reused);

OpenScienceLink

The databases developed during the course of this project could be protected by the *sui generis* database right. This would be the case if the 'right-holder' has made a substantial investment when developing these databases.

Research data from other databases

Another important aspect in the context of the OpenScienceLink project is whether using data from another database with research data is allowed, without securing the authorisation of the producer of that database.

As stated before, bare facts are free and cannot be protected by any intellectual property right. The facts can thus be used for new research.

However, copying substantial portions of another database or the whole existing database with research data is not allowed. In that case, one is retrieving and reusing a substantial portion of an existing database which is normally protected by intellectual property rights. This is also the case if a large part of the data in the pre-existing database is copied out after a critical selection has been made.[36] All these actions require the consent of the producer of the database concerned.

OpenScienceLink

When copying and publishing substantial parts of other databases containing research data in the OpenScienceLink database, it is always necessary to ask consent for these actions to the right-holders of these pre-existing databases. The only exception to this principle is when the pre-existing database is not protected by intellectual property rights.

[36] C-304/07, *Directmedia/Albert-Ludwigs-Universität Freiburg,* 9 October 2008

2.4.4.2 Copyright regime on papers for peer-review

A final important aspect that should be addressed at this stage of the project is the copyright regime applicable to the papers or dissertations that users of the platform wouldsend in for peer-review.

As already clarified, copyright regulations in Europe are not fully harmonised, except in the case of software, databases and photographs. In practice, for the papers, this means that the 'originality' requirement (supra) is defined differently according to Member State. This makes matters quite complex, as the minimum amount of creative effort necessary to give rise to copyright protection depends on the context, and on the national copyright regime of the Member State.[37]

As stated before, the levels of originality range from the rather low requirement of 'skill and labour' in the UK, through the medium requirement of the 'individual character and the personal stamp of the author' in France and Belgium, to a demanding requirement of the 'print of the author's personality that rises above average' in Germany. Furthermore, recent judgements of the European Court of Justice seem to suggest that this criterion has been extended to all types of works and hence, that copyright subject matter has finally been harmonized.[38] This however remains to be seen. In practice, it is generally accepted that literary works such as dissertations and monographs, are sufficiently original to fall in the scope of copyright protection.

OpenScienceLink

In theory, not all works are protected by copyright law. Only the content which fulfils the necessary criteria of a particular country falls under the protective shield of the law. Moreover, it is possible that some content is protected in one Member State, but not in another. In practice however, we can assume that literary works such as the works sent in in the OpenScienceLink platform, are protected by copyright.

If copyright rights are attached to the work that is sent for peer-review, then consent is necessary from the *copyright-holder* (supra) of the work to use this work in the context of the OpenScienceLink platform.

2.4.4.3 Conclusion

The issues stemming from the intellectual property rights status of research data are rather complex. One of the reasons for this is the fact that research data as such cannot be protected because they constitute raw facts. Facts are not a protectable subject matter under intellectual property regulations. We have to take into account that the number of legal uncertainties could make the functioning of the OpenScienceLink platform rather difficult in practice.

It will be necessary to design standard contractual agreements (such as terms of use) describing the intellectual property rights status of the database, the research data provided and the provided papers. Such a license could define the input of the data, the allowed use of the work and could also address other legal issues. As stated before, such a license could also solve certain problems related to privacy and data protection regulations.

[37] P.B. Hugenholtz, M. Van Eechoud, S. Van Gompel et al., Recasting of Copyright and Related Rights for the Knowledge Economy, study prepared for the European Commission ETD/2005/IM/D1/95, Amsterdam, November 2006, online available at: http://www.ivir.nl

[38] M. van Eechoud, Along the road to uniformity- diverse readings of the Court of Justice Judgmetns on copyright work, *JIPITEC* 1 –2012, (60) 60-80 and L. Guibault, Licensing research data under Open Access conditions, in D. Beldiman (ed.), *Information and knowledge: 21st century Challenges in Intellectual property and knowledge governance,* Cheltenham, Edward Elgar, upcoming 2013, (2).

OpenScienceLink
In a later stage, a decision will have to be made on the choice of licensing framework (Creative Commons, Open Data Commons, DPL,).

2.4.5 Evolving EU Requirements to Open Access

Official EU documents set a number of open access related requirements, that have legal implication:

- Access to peer-reviewed scientific publications and research data, that is:
  - Free of charge
  - As early as possible in the dissemination process
  - Enable use and re-use of research results
  - Taken into account challenge of intellectual property rights
- Develop licensing systems in accordance with and without prejudice to the applicable copyright legislation and encourage researchers to retain their copyright while granting licenses to publishers
- Clear guidance for researchers on how to comply with open access policies
- Attach metadata to electronic versions of the research output
- Develop a privacy policy
- Take into account trade secrets, national security and legitimate commercial interests
- Allocate responsibility for who is to preserve the scientific information, together with associated financial planning in order to ensure curation and long-term preservation of research results
- Ensure data integrity and authentication
- Certification mechanisms for repositories
- Link publications to underlying data
- Develop new research indicators and bibliometrics encompassing not only scientific publications but also datasets and other types of output from research activity and the individual researcher's performance
- Provide necessary information and infrastructure regarding open access model
  - Barrier for researchers is often lack of information
  - Fear of contractual disagreements with publishers
  - Enforcement open access policy
- Develop systemic reward and recognition mechanism for data sharing
  - Cf. citation mechanisms and measurements of the data citation impact

Dewatripont and co-authors (2006) and other relevant professional literature point to the following additional requirements:

- Guarantee public access to publicly-funded research results shortly after publication
- Raising researcher awareness of journal quality via quality of dissemination (rather than only via citation)

- Open Archives Initiative set up a standard protocol ensuring interoperability between the archives servers: allows metadata to be retrieved from scattered archives and repositories and to aggregate so it can be searched with single query = wider dissemination
- Persistent digital object identifiers should be preferred to URL
- Reference linking
- OpenURL standard.

3 OpenScienceLink Use Cases

3.1 Common Use Cases across Pilots

3.1.1 CUC0.1: Register to the platform

Use Case ID	CUC0.1
Use Case Name	Register to the platform
Purpose	To register the user in the OpenScienceLink platform.
Initiator	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Primary Actor	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Description	The user fills in the requested data and registers to the system in order to obtain a set of credentials and be allowed to use the OpenScienceLink functionalities.
Pre-condition	The user has entered the OpenScienceLink registration page.
Post-condition	The user has registered to the platform and has been provided with his/her credentials for accessing the OpenScienceLink services.
Use Case Functionality	
Sequence	1. The user opens the platform registration page 2. S/he enters the requested data 3. S/he submits the registration form 4. S/he receives the generated credentials for entering the system.
Further Information	
Information Requirements	To avoid spelling errors and resulting inconsistencies, the user could be assisted during this step through auto-complete functionality based on the underlying vocabularies and models.

Table 12: Use Case CUC0.1

3.1.2 CUC0.2: Log in

Use Case ID	CUC0.2
Use Case Name	Log in
Purpose	To authenticate the users entering the OpenScienceLink platform
Initiator	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Primary Actor	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority

Description	The user enters the OpenScienceLink platform by providing their credentials.
Pre-condition	The user has registered to the OpenScienceLink platform and keeps her/his credentials.
Post-condition	Success end condition The user has been authenticated and can access the OpenScienceLink platform with different privileges depending on their role and authorisation details. Failure end condition: The user has not been allowed to enter the system due to provision of invalid credentials.
Use Case Functionality	
Sequence	1. The user enters their credentials 2. The user enters the OpenScienceLink platform
Exceptions	1a. In step 1 of the normal sequence, if the user enters invalid credentials (either incorrect or non-existent), then an appropriate error message will be prompted to her/him.
Use Cases Dependencies	CUC0.1
Further Information	
Non-functional Requirements	**Performance**: The login process should last no longer than a few msec.

Table 13: Use Case CUC0.2

3.1.3 CUC0.3: Log out

Use Case ID	CUC0.3
Use Case Name	Log out
Purpose	To disconnect the users who have logged in the platform
Initiator	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Primary Actor	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Description	The user presses the log out button in the OpenScienceLink platform.
Pre-condition	The user has logged in the OpenScienceLink platform.
Post-condition	The user has disconnected from the platform.
Use Case Functionality	

Sequence	1. The user presses the log out button 2. The user is disconnected from the OpenScienceLink platform and is prompted with the respective message.
Exceptions	1a. In step 1 of the normal sequence, if the user has logged in and is inactive for more than 20 minutes, then the platform automatically disconnects him/her for security reasons.
Use Cases Dependencies	CUC0.2
Further Information	
Non-functional Requirements	**Performance**: The log out process should last no longer than a few msec.

Table 14: Use Case CUC0.3

3.2 Pilot 1: Research Dynamics-aware Open Access Data Journals Development

3.2.1 UC1.1: Journal Issue Initialization

Use Case ID	UC1.1
Use Case Name	Journal Issue Initialization
Purpose	To initialize the process for the creation of a new Open Access Data Journal Issue. Fields of interest and respective call for datasets should be initiated.
Initiator	Publisher
Primary Actor	Publisher
Description	The publisher logs in to the OpenScienceLink Platform in order to identify scientifically hot topics for the creation of a new journal issue. The platform suggests possible topics for the new journal issue, based on the trend detection and analysis that it performs on that field. The topics are presented in the form of a list, with scores that quantify the trend over time. Alternatively, the publisher directly enters a specific topic, requests for and is presented with its trend analysis measurements. In parallel, the publisher is also able to see from the pool of datasets that have been submitted, which datasets are related to the designated topic. This way the publisher is aware of: (i) the significance of the issue to be released, based on the trend detection and analysis of topics, and, (ii) the coverage that these topics have in the uploaded/submitted datasets by the researchers. Once the topics for the issue are identified, the publisher launches the call for datasets which is from that moment on visible to all of the platform users. The call could also be distributed automatically with e-mail in known mailing lists of scientific interest to the topics of life sciences covered by OSL.
Pre-condition	The publisher is logged in at the OpenScienceLink platform.

Post-condition	A call for data sets initializing an issue for the open access data journal is clearly defined and publicized to the platform and respective mailing lists.
Use Case Functionality	
Sequence	1. Publisher switches to journal issue creation view of the platform (available only to publisher role). 2. Publisher is presented with topics and respective trend score. 3. In parallel to (2) publisher is presented with submitted and unpublished datasets that exist in the platform, with respective annotation of topics. 4. Publisher specifies list of topics for the call for data sets of the new issue. 5. Call is publicized to the platform and other fora automatically through the platform.
Use Cases Dependencies	UC3.1
Further Information	
Assumptions	The publisher is briefly aware of interest in specific topics. He gets, however, support from the platform regarding trends and impact.
Open Issues	Distributing the call to proper media automatically through the platform.
Information Requirements	The list of the topics with their trend scores.
Non-functional Requirements	The call should pertain ideally to topics for which there are already some datasets uploaded. This way there will be at least some datasets for the issue, independently on whether there are new data set submissions for this call.

Table 15: Use Case UC1.1

3.2.2 UC1.2: Data set submission

Use Case ID	UC1.2
Use Case Name	Data set submission
Purpose	To allow researchers submit their data sets to the OpenScienceLink platform for publication/inclusion in future issues of the open access data journal.
Initiator	Researcher
Primary Actor	Researcher

Description	A registered researcher, that has a profile with the OpenScienceLink platform, logs in, in order to submit a data set. The researcher may submit a dataset pertaining to any topic covered by OpenScienceLink at any given time, or assign the data set to be submitted to a specific open call for data sets available through the platform. In this latter case the submitted dataset will be published once the relevant issue or volume of the data journal is released. By using an appropriate user interface, the researcher specifies the dataset's parameters and metadata (e.g., creator, date or period of creation, overall description, purpose, conditions, features or parameters measured, references to related works that used this dataset, how the data set could be utilized, other notes or remarks that are important to bring into the attention of future users of the dataset), and uploads the dataset to the platform, where it is stored for further evaluation. The dataset format will follow the general guidelines specified by the Unidata and Triad data models, and the file may be in any file format. Once the dataset is uploaded it becomes indexed and searchable from all platform users, but not downloadable, which requires that the dataset is first reviewed and then published by the data journal.
Pre-condition	The researcher has an account (profile) created with the OpenScienceLink platform and is logged in.
Post-condition	A data set which is uploaded to the OpenScienceLink platform and indexed.
Use Case Functionality	
Sequence	1. Researcher switches to the data set submission view (available only to researcher role). 2. Researcher is presented with a GUI that allows him to fill all the necessary meta-data information about the data set manually. 3. Researcher assigns also topics to the data set and (optionally) assigns the data set on queue to be published under a specific issue for which a call is at that time active and open. 4. Researcher uploads the actual data set file, and if successful, researcher receives a confirmation message of successful upload.
Further Information	
Open Issues	Defining the exact restrictions regarding the file format.
Information Requirements	The list of supported topics from the platform.

Table 16: Use Case UC1.2

3.2.3 UC1.3: Data set peer review

Use Case ID	UC1.3
Use Case Name	Data set peer review
Purpose	To assign submitted data sets for peer-reviewing and monitor the peer-reviewing process.
Initiator	Publisher/Editor

Primary Actor	Publisher
Additional Actors	Reviewer (Researcher)
Description	The submitted datasets must undergo peer reviewing, in order to be evaluated in terms of a series of factors, such as rationale, detail level, possible impact, completeness, etc. for inclusion in the journal issue. For this purpose, the dataset is primarily connected with related literature, topics, and authors of the life sciences domains, using semantic enabled technologies, such as text annotation with ontology concepts, The annotation is conducted just after the submission of a data set, automatically by the OpenScienceLink platform, considering and analyzing the metadata filled by the researcher who submitted the data set. Through the existing annotations, the publisher can then identify the appropriate reviewers and then may assign certain datasets to certain reviewers for peer-reviewing. The reviewers evaluate the datasets according to different criteria, aided by the tools provided by the platform, such as evaluation forms with criteria, scorings and explanations. When the review is finished, the platform publicizes the reviewers' comments, which can be viewed and discussed upon by other platform users. Once the review of the dataset is over, and the dataset is accepted for publication, it enters into a special status within the dataset pool, flagged as 'ready to be published', and enables the publisher to select it as part of any future volume or issue for publication in the data journal, or for the designated journal by the researcher during the uploading process.
Pre-condition	The publisher and the reviewers (researchers) have accounts with the OpenScienceLink platform. The reviewers can create accounts upon being invited by the platform to review the data set.
Post-condition	A reviewed data set, queued for publishing with the open access data journal.
Use Case Functionality	
Sequence	1. Publisher switches to the data set reviewing view (available only to publisher role). 2. Publisher is able to view the submitted data sets which have not been reviewed yet. 3. For each such data set, suggestions on related expert researchers are given. 4. The publisher selects a list of reviewers for each of the data sets. The platform sends notifications/invitations to all of them. 5. Reviewers log in (existing profile, or create new) and review the data set through an available review form presented by the platform. 6. Reviewers submit the reviews. Reviews become available for discussion. 7. Upon data set acceptance, data sets are queued for publication under the open access data journal.
Use Cases Dependencies	UC1.2
Further Information	
Open Issues	Definition of the evaluation criteria for the data sets.

Information Requirements	List of submitted, but unreviewed data sets, with accompanying annotations on related topics, papers and authors that are experts in the field.

Table 17: Use Case UC1.3

3.3 Pilot 2: Novel open, semantically-assisted peer review process

3.3.1 UC2.1: Reviewers Suggestion and Selection

Use Case ID	UC2.1
Use Case Name	Reviewers Suggestion and Selection
Purpose	To suggest expert reviewers for peer-review of research papers and datasets, and allow their selection and order their invitation to the peer-review process.
Initiator	Publisher/Editor
Primary Actor	Publisher/Editor
Additional Actors	Researcher
Description	The publisher/editor is able to view a list of suggested reviewers for a submitted dataset, or for a submitted paper, and is able to select from the recommendations the list of the reviewers that the system should invite for the peer-review process. The system automatically invites the selected reviewers via the platform to submit the reviews, allowing them to create an account if such does not yet exist, or log-in and access the peer-review forms.
Pre-condition	The user has logged-in the OpenScienceLink platform.
Post-condition	The user has selected a set of reviewers that will perform the peer-review process of the submitted research paper or dataset. The candidate reviewers have been invited automatically via the platform.
Use Case Functionality	
Sequence	1. Publisher switches to the view of uploading papers for review. 2. Publisher uploads the paper to be reviewed in a pdf format. 3. System reads and analyses the paper in reasonable time, and presents the publisher a list of candidate reviewers to perform the review. 4. Publisher selects a number of reviewers among the candidates in the list. Preferably two or more reviewers are selected. 5. Publisher presses the respective button that does the assignment and invites the candidate reviewers for the review. 6. Candidate reviewers are invited via e-mail automatically by the platform.
Alternatives	Especially for the dataset peer-reviewing, the UC1.3 may be followed as an alternative.

Use Cases Dependencies	CUC0.2, UC1.3
Further Information	
Assumptions	The uploaded pdf by the publisher is not corrupted as file. The system acknowledges that this is a paper under review, that is not supposed to be circulated publicly until the reviews are made, unless the publisher wishes otherwise.
Open Issues	The specific fields of the review form. The existence of an option that will allow the publisher/editor to leave the article after the review on the platform for public viewing.
Information Requirements	E-mails of suggested reviewers.
Non-functional Requirements	The publisher/editor consents to upload the pdf of the article at the OSL platform for the purposes of the reviewing and consents to allow the OSL platform to distribute it to the selected reviewers of his choice.

Table 18: Use Case UC2.1

3.3.2 UC2.2: Assisted Peer-Review Submission

Use Case ID	UC2.2
Use Case Name	Assisted Peer-Review Submission
Purpose	To assist the invited reviewers in submitting their review for a submitted article.
Initiator	Publisher/Editor
Primary Actor	Researcher
Description	The researcher has been invited through the OSL platform via e-mail after he has been designated by the Publisher/Editor as one of the reviewers of a submitted article. The researcher logs in to the OSL platform, or creates a new account, and reviews the article with the assistance of the platform. The platform suggests related articles, datasets, topics and top-authors in the respective article fields. The researcher uses this information to fill-in a review form, and submits the review form to the OSL platform with his final decision.
Pre-condition	The user has accepted a review invitation from the OSL platform, and has logged-in to the platform.
Post-condition	The user has submitted a review form for the article he was invited to review.
Use Case Functionality	

Sequence	1. Researcher accepts an invitation sent via e-mail from the OSL platform, to review a submitted article. 2. Researcher logs-in to the platform, or creates a new account if one does not already exist. 3. Researcher is presented with the pdf of the submitted paper and an online electronic review form. 4. Researcher is automatically assisted by the platform by being presented with related articles, datasets, topics and top authors in the submitted article research field(s). 5. Researcher makes use of the platform information, to fill in the review form. 6. Researcher decides on the outcome of the review for the article (major part of the review form) and submits the review form to the OSL platform by clicking a 'submit' button.
Alternatives	Especially for the dataset peer-reviewing, the UC1.3 may be followed as an alternative.
Use Cases Dependencies	CUC0.2, UC1.3
Further Information	
Assumptions	The uploaded pdf by the publisher is not corrupted as file. The system acknowledges that this is a paper under review, that is not supposed to be circulated publicly until the reviews are made, unless the publisher wishes otherwise.
Open Issues	The specific fields of the review form. The existence of an option that will allow the publisher/editor to leave the article after the review on the platform for public viewing.
Information Requirements	Meta-data of the article (author names), and/or related research fields that will provide input to the platform to collect the assisting material for the reviewer.
Non-functional Requirements	The publisher/editor consents to upload the pdf of the article at the OSL platform for the purposes of the reviewing and consents to allow the OSL platform to distribute it to the selected reviewers of his choice.

Table 19: Use Case UC2.2

3.3.3 UC2.3: Final Review Decision

Use Case ID	UC2.3
Use Case Name	Final Review Decision
Purpose	To draw a final decision on the review of a submitted article.
Initiator	Publisher/Editor
Primary Actor	Publisher/Editor
Additional Actors	Researcher

Description	The Publisher/Editor has been informed by the OSL platform via e-mail that all of the assigned reviews for a submitted paper are now deposited in the OSL platform. The Publisher/Editor logs in to the platform, browses the submitted review forms and draws a final decision on the outcome of the review. If the Publisher/Editor wishes so, the final decision is communicated to the assigned reviewers via e-mail through the OSL platform. The Publisher/Editor also decides whether he wants to leave the submitted article with the OSL platform to be indexed for public viewing (optionally along with the reviews).
Pre-condition	The assigned reviewers (researchers) for a submitted article have submitted their review forms for this article.
Post-condition	A decision is drawn for the reviewing of a submitted article.
Use Case Functionality	
Sequence	1. Publisher/Editor receives a notification via e-mail from the OSL platform that all of the assigned reviewers for a submitted article have deposited their review forms with the OSL platform. 2. Publisher/Editor logs in to the platform and browses the submitted review forms. 3. Publisher/Editor draws a decision based on the reviewers' recommendations. 4. Optionally, Publisher/Editor allows the OSL platform to communicate the decision to the reviewers. 5. Optionally, Publisher/Editor allows the OSL platform to index the submitted pdf (perhaps with the reviewers' forms) for public viewing and search from the OSL platform users.
Use Cases Dependencies	CUC0.2, UC1.3, UC2.2
Further Information	
Assumptions	The Publisher/Editor is able to draw a decision on the review outcome of the submitted article by consulting the submitted review forms.
Open Issues	Allowing the reviewer comments to be publicly available in case the Publisher/Editor wishes to leave the pdf for indexing with the OSL platform.

Table 20: Use Case UC2.3

3.4 Pilot 3: Data mining for Biomedical and Clinical Research Trends Detection and Analysis

3.4.1 UC3.1: Detection of Research Trends

Use Case ID	UC3.1
Use Case Name	Detection of Research Trends
Purpose	To provide an analysis and respective visualization of research trends in the life science fields of research.
Initiator	Researcher

Primary Actor	Researcher
Description	The user can obtain a visualization of the research trends and hot topics in the life science fields of research. The visualization shows the trends of the most important hot topics based on the analysis conducted by examining the published literature topics in time slots, and computing the topics acceleration in terms of frequency of appearance.
Pre-condition	The user has logged-in the OpenScienceLink platform.
Post-condition	The user sees a visualization of the trends in the research fields of the life sciences.
Use Case Functionality	
Sequence	1. The user switches to the trends analysis view of the OSL platform. 2. The user views plots that take into account the temporal dimension, of the frequency of appearance and acceleration (trend) for all of the fields of the life sciences. The fields are ordered by default in a descending manner with regards to the computed trends. 3. The user selects specific fields to zoom in and get more detailed analysis on hot topics in that research field. 4. The user can also alter the temporal parameters of the visualization to zoom in and out in specific time spans.
Use Cases Dependencies	CUC0.2
Further Information	
Assumptions	The trend plots are drawn over time, with a default time unit being used. Potential parameterization of the time unit will be considered.
Open Issues	Under-represented research fields, false positives and false negatives of the research fields tags annotations.
Information Requirements	A hierarchy of life science research fields, taken from the respective branch of the Medical Subject Headings. Annotations of the field tags in the indexed content sources are required to perform the data mining analysis of the trends.

Table 21: Use Case UC3.1

3.4.2 UC3.2: Trends in Authors and Institutions

Use Case ID	UC3.2
Use Case Name	Trends in Authors and Institutions
Purpose	To provide an analysis and respective visualization of trends in authors and institutions in the life sciences scientific literature.
Initiator	Researcher
Primary Actor	Researcher

Description	The user can obtain a visualization of the rising authors and institutions (that experience trend) in the life science literature. The visualization shows the trends of the most important authors and institutions based on the analysis conducted by examining the published literature in time slots, and computing the acceleration in terms of frequency of appearance.
Pre-condition	The user has logged-in the OpenScienceLink platform.
Post-condition	The user sees a visualization of the rising authors and institutions in terms of published work in the research fields of the life sciences.
Use Case Functionality	
Sequence	1. The user switches to the trends analysis view of the OSL platform. 2. The user views plots that take into account the temporal dimension, of the frequency of appearance and acceleration (trend) for all authors and institutions that have published work in life sciences. The authors and institutions are ordered by default in a descending manner with regards to the computed trends. 3. The user selects specific authors or institutions to zoom in and get more detailed analysis on their trend. 4. The user can also alter the temporal parameters of the visualization to zoom in and out in specific time spans.
Use Cases Dependencies	CUC0.2
Further Information	
Assumptions	The trend plots are drawn over time, with a default time unit being used. Potential parameterization of the time unit will be considered.
Open Issues	Under-represented authors, availability of only lead/contact institutions for the institution analysis. Disambiguation of authors and institutions.
Information Requirements	A list of authors and institutions that have published work within any field of life sciences (obtained by GoPubMed).

Table 22: Use Case UC3.2

3.5 Pilot 4: Data mining for proactive formulation of scientific collaborations

3.5.1 UC4.1: Request for Collaboration Suggestions

Use Case ID	UC4.1
Use Case Name	Request for Collaboration Suggestions
Purpose	To submit a request for being presented with suggestions of potential scientific collaborations with researchers, research groups and/or communities.
Initiator	Researcher, Publisher
Primary Actor	Researcher, Publisher
Additional Actors	
Description	The user can request for suggestions of scientific collaborations with researchers, research groups and/or communities.
Pre-condition	The user has logged-in the OpenScienceLink platform.
Post-condition	The user has submitted a request for retrieving a list of researchers, research groups and/or communities with which they could potentially collaborate.
Use Case Functionality	
Sequence	1. The user enters the scientific collaboration part of the platform 2. S/he chooses the collaboration target, i.e., Researcher, Research Group and/or Research Community 3. (optional) S/he chooses the sources to be used for generating the collaborations to be suggested 4. (optional) S/he chooses to receive regular notifications about proposed research collaborations via e-mail
Use Cases Dependencies	CUC0.2

Table 23: Use Case UC4.1

3.5.2 UC4.2: View Collaborations Suggestions

Use Case ID	UC4.2
Use Case Name	View Collaboration Suggestions
Purpose	To view the generated suggestions of scientific collaborations with researchers, research groups and/or communities.
Initiator	Researcher, Publisher
Primary Actor	Researcher, Publisher
Description	The user (researcher or publisher) can view the generated suggestions of scientific collaborations with researchers in the form of a list of names.
Pre-condition	The user has logged-in the OpenScienceLink platform and has submitted a request for retrieving suggestions of collaborations with researchers, research groups and/or communities.

Post-condition	The user has gone through the suggestions for collaborations and may or may not choose to initiate an effort for collaboration with one or more of the suggested researchers, research groups and/or communities.
Use Case Functionality	
Sequence	1. The user is prompted with the list of researchers, research groups and/or communities with which OpenScienceLink suggests s/he could collaborate with 2. Each of the researchers' names is accompanied by a subset of their scientific papers which are highly related to the user's research work 3. (optional) For each proposed collaboration, the user is presented with a percentage indicating how strong the suggested collaboration is.
Alternatives	2a. In step 2 above, if the user is a publisher, then the list of suggested researchers depends on their scientific work's relevance with the journal(s) and/or journal issue(s) they are responsible for. 2b. In step 2 above, if the collaboration target is research groups and/or communities, then the list of the suggested ones depends on their profile (as exposed through their metadata) and their members profiles. 2c. In step 2 above, if the collaboration target is research groups and/or communities and the user is a researcher, then the list of the suggested ones is accompanied by the scientific topics, fields and/or areas that they cover and are common with the researcher's interest. 2d. In step 2 above, if the collaboration target is research groups and/or communities and the user is a publisher, then the list of the suggested ones for dissemination purposes is accompanied by the scientific topics, fields and/or areas that they cover and are common with the publisher's journal(s) and/or journal issue(s) scientific focus.
Use Cases Dependencies	CUC0.2, UC4.1

Table 24: Use Case UC4.2

3.5.3 UC4.3: Receive Notification for Suggestions of Collaboration

Use Case ID	UC4.3
Use Case Name	Receive Notification for Suggestions of Collaborations with Researchers
Purpose	To view the generated suggestions of scientific collaborations with researchers.
Initiator	Researcher, Publisher
Primary Actor	Researcher, Publisher
Description	The user (researcher or publisher) receives regular updates for suggested collaborations with researchers, research groups and/or communities via e-mail.

Pre-condition	The user has submitted a request for receiving regular notifications for collaboration suggestions.
Post-condition	The user has received an e-mail with a list of suggested collaborations and may or may not choose to initiate an effort for collaboration with one or more of the suggested researchers, research groups and/or communities.
Use Case Functionality	
Sequence	1. The user receives an e-mail which includes a list of researchers with whom OpenScienceLink suggests s/he could collaborate with 2. The user is suggested to follow an accompanying link to the OpenScienceLink portal for viewing more information about the suggested collaborations, including a subset of their scientific papers which are highly related to the user's research work 3. (optional) For each proposed collaboration, the user is presented with a percentage indicating how strong the suggested collaboration is.
Alternatives	2a. In step 2 above, if the user is a publisher, then the list of suggested researchers depends on their scientific work's relevance with the journal(s) and/or journal issue(s) they are responsible for.
Use Cases Dependencies	UC4.1

Table 25: Use Case UC4.3

3.5.4 UC4.4: Receive Suggestion for Participation in Research Community

Use Case ID	UC4.4
Use Case Name	Receive Suggestion for Participation in Research Community
Purpose	To inform the user about new, automatically-formed research communities generated by OpenScienceLink.
Initiator	Researcher, Publisher
Primary Actor	Researcher, Publisher
Description	The user is presented with suggestions regarding absolutely related research communities and entities to his research profile.
Pre-condition	The user has logged in the OpenScienceLink platform.
Post-condition	The user has been prompted with the option to join a research community being automatically and dynamically formed by the OpenScienceLink.
Use Case Functionality	

Sequence	1. The platform correlates the user's profile (in the case of the researcher, their research topics, field and/or domain of the researcher and in the case of the publisher, the journal's scientific focus) with the platform-generated research communities 2. The user is presented with a suggestion for joining one or more research communities which are being formulated currently or recently have been established by the OpenScienceLink and which best fit their profile 3. The user chooses to join at least one of them
Alternatives	3a. In step 3 above, if the user is not interested in participating in any of the suggested communities, then s/he may choose none of them. In this case, s/he is prompted with a question asking him the reason for not joining any of them (in order for the platform to capture whether the suggestions were of low relevance to the user's work)
Use Cases Dependencies	CUC0.2

Table 26: Use Case UC4.4

3.5.5 UC4.5: Receive Suggestion for Leading a Research Community

Use Case ID	UC4.5
Use Case Name	Receive Suggestion for Leading a Research Community
Purpose	To inform the user about new, automatically-formed research communities generated by OpenScienceLink and propose his leadership based on their excellence in the topic/field/domain.
Initiator	Researcher
Primary Actor	Researcher
Description	The user is presented with suggestions with regards to which research fields he appears to be among the lead authors/researchers.
Pre-condition	The user has logged-in to the OpenScienceLink platform.
Post-condition	The user is presented with a list of research fields where he appears to be among the lead authors/researchers.
Use Case Functionality	
Sequence	1. The platform correlates the user's profile (in the case of the researcher, their research topics, field and/or domain of the researcher and in the case of the publisher, the journal's scientific focus) with the platform-generated research communities 2. If the user is among the top researchers, based on the evaluation of the user in the topic(s), field(s) and/or domain(s) that these research communities focus on, the user is presented with a suggestion for leading one or more of these communities 3. The user choose to join at least one of them

Alternatives	3a. In step 3 above, if the user is not interested in leading in any of the suggested communities, then s/he may choose none of them. In this case, s/he is prompted with a question asking him the reason for not joining any of them (in order for the platform to capture whether the suggestions were of low relevance to the user's work)
Use Cases Dependencies	CUC0.2

Table 27: Use Case UC4.5

3.6 Pilot 5: Scientific Field-aware, Productivity and Impact-oriented Enhanced Research Evaluation Services

3.6.1 UC5.1: Retrieve Evaluation for a Specific Object

Use Case ID	UC5.1
Use Case Name	Retrieve Evaluation for a Specific Object
Purpose	To allow the user to have access to the evaluation of a researcher, journal, paper and/or data set s/he chooses
Initiator	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Primary Actor	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Description	The user views the evaluation of a paper, data set, journal, researcher, research group, community, institution, country
Pre-condition	The user has logged in the OpenScienceLink platform.
Post-condition	The user has viewed the requested evaluation.
Use Case Functionality	
Sequence	1. The user selects the type of the evaluation object (i.e., paper, data set, journal, researcher, research group, community, institution, country). 2. (optional) S/he chooses a research topic, field and/or domain within which s/he prefers the evaluation to be narrowed 3. The user enters the evaluation object's details (e.g., the researcher's name, the title of the paper, etc) 4. The user is presented with the available evaluation metrics 5. (optional) S/he chooses based on which s/he prefers to view evaluation results in each case (paper, data set, journal, etc). 6. S/he views the evaluation results, which cover individual evaluation metrics and the overall evaluation.
Use Cases Dependencies	CUC0.2

Table 28: Use Case UC5.1

3.6.2 UC5.2: Retrieve Top Representatives in a Specific Topic

Use Case ID	UC5.2
Use Case Name	Retrieve Top Representatives in a Specific Topic
Purpose	To allow the user to have access to the evaluation of a researcher, journal, paper and/or data set s/he chooses
Initiator	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Primary Actor	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Description	The user is able to view the evaluation of the top subjects and objects (i.e., researchers, papers, data sets, etc) in a specific research topic, field or area.
Pre-condition	The user has logged in the OpenScienceLink platform.
Post-condition	
Use Case Functionality	
Sequence	1. S/he chooses a research topic, field and/or domain within which s/he prefers the evaluation to be narrowed. 2. (optional) The user selects the type of the evaluation object (i.e., paper, data set, journal, researcher, research group, community, institution, country). 3. (optional) S/he chooses the evaluation metrics based on which s/he prefers to view evaluation results in each case (paper, data set, journal, etc). 4. The user is presented with the top subjects and objects (i.e., researchers, papers, data sets, etc). 5. For each one of them, s/he can view the individual evaluation metrics and the overall evaluation. 6. (optional) S/he can rank the results based on each evaluation metric as well as the overall evaluation.
Use Cases Dependencies	CUC0.2

Table 29: Use Case UC5.2

3.6.3 UC5.3: Retrieve Own Evaluation

Use Case ID	UC5.3
Use Case Name	Retrieve Own Evaluation

Purpose	To allow the user to have access to their own evaluation based on their profile (i.e., researcher or publisher)
Initiator	Researcher, Publisher
Primary Actor	Researcher, Publisher
Description	The user accesses the evaluation related to their work
Pre-condition	The user has logged in the OpenScienceLink platform.
Post-condition	The user views evaluation metrics and indices for his own research.
Use Case Functionality	
Sequence	1. The user chooses to view his/her own evaluation. 2. (optional) The user selects the type of the evaluation object s/he is interested in (i.e., paper, data set, research group, institution). 3. (optional) S/he chooses the evaluation metrics based on which s/he prefers to view evaluation results in each case. 4. The user is presented with the requested evaluation results, which cover individual evaluation metrics and the overall evaluation. 5. (optional) S/he can rank the evaluation objects based on each evaluation metric as well as the overall evaluation.
Alternatives	2a. In step 2 above, if the user is a publisher, then the available types of evaluation objects include journal and journal issue (and optionally papers and data sets).
Use Cases Dependencies	CUC0.2

Table 30: Use Case UC5.3

3.6.4 UC5.4: Follow Evaluations

Use Case ID	UC5.4
Use Case Name	Follow Evaluations
Purpose	To allow the user to follow evaluations of papers, data sets, journals, researchers, research groups, communities, institutions and countries they are interested in
Initiator	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Primary Actor	Researcher, Publisher, Member of Research Group, Community, Academic and/or Research Institute, Representative of Funding Agency or Authority
Description	The user selects the paper(s), data set(s), journal(s), researcher(s), research group(s), communities, institution(s) and countries for which s/he would like to follow their evaluation and is provided with prompt access to their evaluation results.
Pre-condition	The user has logged in the OpenScienceLink platform.

Post-condition	The user is presented with evaluation metrics and indices with regards to the selected entities (paper(s), data set(s), journal(s), researcher(s), research group(s), communities, institution(s) and country(ies)).
Use Case Functionality	
Sequence	1. The user chooses the paper(s), data set(s), journal(s), researcher(s), research group(s), communities, institution(s) and countries s/he would like to follow. This option is provided each time they view their search results and/or as a separate menu option. 2. (optional) S/he chooses the evaluation metrics based on which s/he prefers to view evaluation results in each case. 3. The user is presented with the requested evaluation results, which cover individual evaluation metrics and the overall evaluation. 4. (optional) S/he can rank the evaluation objects based on each evaluation metric as well as the overall evaluation.
Use Cases Dependencies	CUC0.2

Table 31: Use Case UC5.4

4 OpenScienceLink Content Sources

4.1 OpenScienceLink Requirements for Content Sources

In this chapter, we describe the different models and initiatives on sharing biomedical data to demonstrate best practices and lessons learned. With the variety of open data initiatives presented it gives a thorough view on the types of data sources available, the spectrum of open data research projects and consortiums, the technical and cultural challenges faced, as well as some unresolved issues. Given the selection of clinical and biological research as the validation domain for the project's pilot services, the project will leverage a wide range of openly accessible repositories, which will empower the credible piloting and evaluation of the project's platform and overall approach to open access. In particular, the project will exploit the following repositories of scientific information:

- The Directory of Open Access Journals (http://www.doaj.org/), providing access to more than 7,400 open access journals.
- PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), at abstract level, comprising more than 23 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed citations include links to full-text content from PubMed Central and publisher web sites. In addition, approximately 3 million articles of the aforementioned number, are open access journal articles.
- ClinicalTrials.gov, comprising information of over 125,000 clinical trials including design, status and results.
- Clinicaltrialsregister.eu, containing information about more than 16,000 EU Clinical Trials.
- Citeseer, one of the most prominent public search engines and digital library for scientific and academic papers, primarily in the fields of computer and information science.
- The open access journals of OMICS Group (see: http://www.omicsonline.org/open-access-journals-list.php/), which are approximately 500 journals pertaining to the biomedical domain.
- DrugBank, containing information about more than 6,700 drugs.
- The LinkedData source Diseasome, which provides access to more than 500 diseases and genes and more than 1,000 relationships among them.
- The LinkedData source SIDER, which provides access to more than 1,500 side effects, more than 880 drugs and more than 62,000 relationships among them
- The LinkedData source KEGG, which provides access to more than 9,000 drugs, more than 16,000 molecules, more than 7,1 million entries for genes and more than 64,225 billion amino acid sequence similarities among all protein-coding genes in the complete genomes, among others
- The IUCLID Chemical Data Sheets Information System, comprising Chemical Data Sheets for over 2600 substances.
- The UniProt database, which provides data about 500,000 reviewed protein sequences (approximately the number of the SwissProt sequences inside UniProt), and 23 unreviewed protein sequences (approximately the number of the TrEMBL sequences inside UniProt). Almost all of the SwissProt sequences can be found/annotated in the current biomedical bibliography, while linking the 23 million unreviewed sequences to the bibliography will be a challenge, but also a great test case for the linking of biomedical data with scientific publications that OpenScienceLink should be able to provide.
- The RCSB PDB (Protein Data Bank), which contains information about 78,760 proteins, 2,431 nucleid acids, and 3,843 protein complexes. The information pertains to the 3-D structure of the sequences, as well as to the experimental method with which the sequences were identified (e.g., X-RAY, NMR, Electron Microscopy, Hybrid).

- Scientific predictions, which are not yet published, regarding the interactions between biomedical entities extracted from both full text closed journal articles and full text open access journal articles in the biomedical domain (estimated at approximately 500 million predictions).
- Protein-to-protein interaction data (predictions), stemming from text mining methodologies, e.g., ontological profile, estimated to approximately 5,000 predicted interactions with high prediction score.

These repositories are related and will be integrated and linked in the scope of the OpenScienceLink platform. Access to most of these repositories is free and without any limitations associated with IPR issues. However, the project will (as part of its work plan) investigate any IPR implications or limitations prior to the use of the data sources in the project, since this will be also a task associated with the implementation and integration of the legal framework and tools of the OpenScienceLink of the platform.

The above data repositories ensure a critical mass of data for testing, validating and evaluating the pilot services of the project. This is because they can support end-to-end processes associated with the pilot services such as end-to-end review processes, computations of scientometrics and development and linking of data journals. At the same time, the project will rely on the generation and publishing of open data from the users (e.g., researchers, scholars) that will participate in the OpenScienceLink pilots. The capabilities of the project's platform will instigate the development of data journals, as well as their linking with data and information contained in the above-mentioned data repositories. Furthermore, the project will take additional measures towards populating its platform with a critical mass of data needed for validation, in particular:

- The project partners will make available (as openly accessible information) research articles from readily available collections. This applies to both the research/academic partners (i.e. TUD, NKUA, CNR, LUHS) and the SMEs of the consortium (i.e., PROCON and TI). Note that IPR issues will be investigated for all the additional data sets and data repositories to be included in the project.
- The establishment of a stakeholders/users group comprising third-party organizations (i.e., outside the consortium), which will engage in the project as end-users, while also contributing content repositories. The establishment and gradual expansion of this stakeholders' group will be part of the project's dissemination and exploitation plans.

Finally, with regards to the scientific data that will be integrated to the platform, the aforementioned list covers in a satisfactory manner the research data relating to experiments that are not necessarily published in peer-reviewed papers, but it does not cover in a large degree the research data underlying open/closed access scientific publications (final datasets and/or raw data). In this direction, in the following section we review additional content sources that can potentially address this need, which are either data repositories, or projects and/or initiatives through which actual scientific data are made available.

4.2 Evaluated Data Content Sources

OPEN DATA INITIATIVES AND MODELS OF DATA SHARING IN THE BIOMEDICAL DOMAIN

ClinicalTrials.gov

http://www.nlm.nih.gov/pubs/factsheets/clintrial.html

Establishment in the year 2000, ClinicalTrials.gov serves as a registry of clinical trials at the trials' inception (Zarin, et al., 2011). The registry now contains key protocol details of more than 130,000 trials from around the world. In 2008 the registry added a results database, which now contains the summary results of more than 7,000 trials.

Clinical trials data take many forms, from uncoded, participant-level data to analyzed summary data; only the latter are posted at ClinicalTrials.gov. There are three key problems with the practice of evidence-based medicine. Not all trials are published. As Deborah Zarin stated, publications do not always include all of the prespecified outcome measures, while unacknowledged changes made to trial protocols can affect the interpretation of findings (Zarin, 2012).

At each step in the process leading from the raw data to the summary data, information is lost. Also, each step involves subjective judgments that are not transparent, but can influence the reproducibility of results. The users of summary data generally assume that they reflect the underlying participant-level data, with little room for subjectivity. That assumption is not always correct.

The results database at ClinicalTrials.gov was launched in response to the Food and Drug Administration Amendments Act of 2007 and was based on statutory language and other relevant reporting standards. It requires that the sponsors or investigators of trials report the "minimum dataset," which is the dataset specified in the trial protocol in the registry. The data are presented in a tabular format with minimal narrative. They cover participant flows, baseline patient characteristics, outcome measures, and adverse events. The European Medicines Agency is currently developing a similar results database. While ClinicalTrials.gov has checks for logic and internal consistencies, it has no way of ensuring the accuracy of the data reported.

ClinicalTrials.gov does not dictate how data are analyzed, but does require that the reported data make sense. ClinicalTrials.gov was established on the assumption that required data are generated routinely after a clinical trial based on the protocol for the trial, so the burden of reporting to ClinicalTrials.gov would be due mainly to data entry. Instead, the experience at ClinicalTrials.gov has shown that protocols are often vague, are not always followed, or in some cases may not even exist. In addition, summary data are not always readily available even for trials that have already been published. For many trials, no one can explain the structure of the trial or the analysis of the data. According to Zarin, "there is not an objective, easy-to-describe route from the initial participant-level data to the summary data. Many people and many judgments are involved" (Zarin, 2012).

Structural changes to trials are also common. A trial can start as a two-arm study and then become a four-arm study. Participants come and go, so that the number of participants changes over time. Participant flow and baseline characteristic tables describe different populations than

the outcomes table. Data providers often cannot explain the "denominators" for their results, the groups from which outcomes or adverse events are collected. In other cases, outcome measures were changed: a quality-of-life scale was replaced with a depression scale; 1-month data were replaced with 3-month data; the number of people with an event was replaced with time to an event; and all-cause mortality was replaced with time to relapse. Sometimes discrepancies are obvious. In one study, the mean hours of sleep per day was listed as 823.32 hours. Another study of 14 people included data on 36 eyeballs.

The inevitable conclusion is that summary data may not always be an accurate reflection of participant-level data. Although the deposition of clinical trial protocols and summary data into registries is a huge step forward in the direction of transparency, the validity and reproducibility of summary data are called into question by such inconsistencies.

Providing more transparency about the process of converting one type of data into another type would help inspire trust. Documents that may help explain this journey include the protocol and amendments, the statistical analysis plan, informed consent forms, clinical study reports, and adverse event reports. Greater transparency would also help everyone involved with clinical trials to engage in internal quality improvements (Zarin, 2012).

DataSphere Project (CEO Roundtable on Cancer)

http://ceo-lsc.org/projectdatasphere

In contrast to the declining mortality rates for heart disease, mortality rates for cancer have dropped only slightly in recent decades. Data sharing in the field of oncology could lead to faster and more effective research through improved trial designs and statistical methodology, the development of secondary hypotheses and enhanced understanding of epidemiology, collaborative model development, and smaller trial sizing. For example, as oncology researchers divide cancers into smaller subgroups with particular molecular drivers, data increasingly need to be pooled to have the statistical power to determine the most effective treatments for each subgroup.[39]

An ideal data-sharing system is simple, systematic, publicly accessible, and respectful of privacy issues. DataSphere, which is an initiative of the CEO Roundtable on Cancer, is designed to achieve these objectives. The CEO Roundtable on Cancer consists of the chief executive officers (CEOs) of companies involved in cancer research and treatment who are seeking to accomplish what no single company can do alone. DataSphere will rely on the convening power of CEOs, together with support from patients and advocacy groups, to secure and provide data. Initially, it will seek to provide comparator arms, genomic data, protocols, case report forms, and data descriptors from industry and academia. DataSphere will include data from both positive and negative studies because a negative study is often as revealing from an epidemiological point of view as a positive study. De-identification will be standardized, and DataSphere will then work with third party data aggregators to pool the data in meaningful ways – a significant challenge when hundreds of cancer drugs are being developed at any given time and thousands of studies are registered in ClinicalTrials.gov.

[39] Presentation of Charles Hugh-Jones, vice president and head of Medical Affairs North America for Sanofi Oncology, at the Workshop on Sharing Clinical Research Data.

DataSphere has established incentives for data contributors that call attention to the increased productivity, cost savings, citations, and collaboration that can accompany sharing. It also is looking at micro-attribution software that could extend credit for sharing to the contributors of data. Similarly, incentives for patients emphasize the benefits of making data available and the security precautions that have been taken. It has even been looking into the possibility of competitions among researchers to enhance the sharing of data.

Tools to enable sharing include a standard de-identification system being developed in collaboration with Vanderbilt University that is consistent with Health Insurance Portability and Accountability Act (HIPAA) regulations, a single online data use agreement form, how-to guides for de-identification, and tools for advocacy. Finally, it has been working closely with the database company SAS to produce a simple but secure, powerful, and scalable website where everything needed to share data is automated.

Sanofi is contributing de-identified data from two recent Phase III clinical studies to start the ball rolling. The goal is to have at least 30 high-quality datasets in the database by the end of 2013 and then expand beyond that.[40]

Sharing Detailed Research Data Is Associated with Increased Citation Rate

A study by Heather A. Piwowar and co-authors examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48 % of trials with publicly available microarray data received 85 % of the aggregate citations. Publicly available data was significantly ($p = 0.006$) associated with a 69 % increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression (Piwowar, et al., 2007). Thus, authors convincingly show that papers for which the data are made available are cited more than papers for which the data are not available.

The Yale-Medtronic Experience

One paradigm for facilitating dissemination of industry data and ensuring high-quality independent review of the evidence for efficacy is exemplified by the Yale-Medtronic experience, where proprietary data were released to an external coordinating organization that contracted other organizations to perform systematic reviews of the study results.[41]

In 2002, the Food and Drug Administration (FDA) approved a product from Medtronic called INFUSE, which was designed to accelerate bone growth in cases of anterolateral lumbar interbody fusion. Approval was based on one pilot randomized controlled study and two pivotal randomized controlled studies. A series of subsequent peer-reviewed publications supported by Medtronic provided additional data on the use of the product.

In June 2011, a major challenge was raised regarding the validity of all the published literature on INFUSE. The principal focus was on the results presented in the peer-reviewed literature and on general study designs and endpoints. The challenge was published in a dedicated issue of a medical journal and consisted of more than 10 articles. The company quickly reviewed its data

[40] Ibid.

[41] Richard Kuntz, senior vice president and chief scientific, clinical, and regulatory officer of Medtronic, Inc., at the Workshop on Sharing Clinical Research Data.

to ensure that the dossiers it had were accurate. The company was convinced that the data were good, and they talked to the FDA immediately to make sure that they are aware of that. However, the issue was being discussed extensively in the media. And the company had to make quick decisions. Within less than a month, the company announced its decision to contract with Yale University as an independent review coordinator. In August, Yale announced its plan to establish an independent steering committee and contract with two systematic review organizations to carry out reviews of the research. Medtronic agreed to supply Yale with all de-identified patient-level data, including non-label studies, along with all FDA correspondence and adverse event reports. It also agreed to allow Yale to establish a public transparency policy and process for the entire INFUSE patient-level dataset. The publication of the systematic reviews was scheduled for the fall and winter of 2012, with summary manuscripts prepared and submitted for publication in the Annals of Internal Medicine.

The project has been undertaken by the Yale University Open Data Access (YODA) project, which serves as a model for the dissemination and independent analysis of clinical trial program data. This project is based on the rationale that a substantial number of clinical trials are conducted but never published, and even among published clinical trials, only a limited portion of the collected data is available. As a result, patients and physicians often make treatment decisions with access to only a fraction of the relevant clinical research data. Clinical trials are conducted with both public and private funding, but several issues are particularly important among industry trials. Industry funds the majority of clinical trial research on drugs, devices, and other products, both pre-market and post-market. Also, industrial research is proprietary, with no requirement for publication or dissemination, and the public perception is that industry has a financial interest in promoting "supportive" research and not publishing the rest of the data.

The YODA project has been designed to promote wider access to clinical trial program data, increase transparency, protect against industry influence, and accelerate the generation of new knowledge. The public has a compelling interest in having the entirety of the data available for independent analysis, but industry has legitimate concerns about the release of data. Steps therefore are needed to align the interests of industry and the public, particularly when concerns about safety or effectiveness arise.[42]

Yale and Medtronic spent a year working through issues involved in assembling the data and giving those data in the most unbiased way possible to reviewers so they could do a full systematic review. To maintain transparency and independence, formal documentation of communications between Yale and Medtronic was necessary along with clarity about what kinds of discussions could and could not be held. The query process among the reviewers, Yale, and Medtronic also had to be carefully managed.

The de-identification process was complicated and expensive. De-identifying the necessary HIPAA fields and information took several months and the efforts of about 25 people, which contributed substantially to the overall $2.5 million cost of the project. The HIPAA Privacy Rule was not designed for this kind of activity. As a result, the YODA project's approach to de-identification was a "Rube Goldberg contraption" and clearly not scalable. Given that paper case report forms and studies going back to 1997 had to be reviewed, the project was "an outlier example of how complicated it would be to deidentify [data]." [43]

Industry has several reasons for participating in this kind of process. It allows fair and objective assessment of product research data, as opposed to speculative analysis based on incomplete

[42] Ibid.
[43] Ibid.

data. It supports competition on the basis of science rather than marketing. It promotes transparency and advances patient care. Although committed to transparency, Medtronic was concerned about potential misuses of the data. In the end, Medtronic sought to provide the data and initiate conversations about its use.

A large number of questions that the Yale-Medtronic project has not fully answered has been raised and are worth considering by anyone who embarks on this kind of research:

- Would it be possible for an independent group to determine whether a question requiring the use of data serves the public interest or a special interest?
- Should queries be limited to single questions, and should the methods used to answer the questions be prespecified?
- Should there be an initial time period during which data remain proprietary?
- What portion and level of the dataset are necessary?
- Should there be a time limit or license for data access?
- Who controls the data distribution?
- Are there a priori questions and hypotheses to be tested, or is there an interest in data exploration?
- Is the requester competent to do the proposed analysis?
- Should a trusted third party analysis center be contracted? May the requester share the data with others?
- Should there be controls on the dissemination of results, such as a requirement for peer review before dissemination?
- What methodological review is required?
- Should industry be involved in the peer review of results derived from its data?

The movement from keeping data concealed to sharing data will require foundational changes. One important step will be involving patients as partners rather than "subjects," which will help lower at least some of the barriers to the use of data.

The Biomarkers Consortium

http://www.biomarkersconsortium.org/press_release_adiponectin_predictive_biomarker.php

The Biomarkers Consortium of the Foundation for the National Institutes of Health (FNIH) is a precompetitive collaboration designed to increase the efficiency of biomarkers-related research. Its goals are to facilitate the development and validation of new biomarkers; help qualify these biomarkers for specific applications in diagnosing disease, predicting therapeutic response, or improving clinical practice; generate information useful to inform regulatory decision making; and make Consortium project results broadly available to the entire scientific community.

Validation of adiponectin as a biomarker is an example of the work of the Consortium. Adiponectin is a protein biomarker discovered in the 1990s that is associated with obesity and insulin sensitivity. Certain drugs can drive up adiponectin levels very quickly in healthy volunteers and in patients, and attention was focused on the use of adiponectin as a predictive biomarker to identify patients who would or would not respond to particular therapies. Though considerable data about adiponectin existed in the files of companies and academic laboratories, relatively few data about the use of adiponectin as a biomarker were publicly available. The Biomarkers Consortium took on the task of compiling these data as a proof-of-concept project for the collaboration. A number of companies agreed to combine their data into a blind dataset derived from many trials involving more than 2,000 patients. Using these data, the consortium

concluded that adiponectin is a robust predictor of glycemic response to peroxisome proliferator–activated receptor agonist drugs used in the treatment of diabetes.

The results confirmed previous findings and investigators concluded that "the potential utility of adiponectin across the spectrum of glucose tolerance was well demonstrated" (Wagner, et al., 2009).

Several important lessons can be drawn from this experience. The project demonstrated that cross-company collaboration was a robust and feasible method for doing this kind of research. However, the project took a relatively long time to complete, which is a real problem. The Consortium has since learned how to collaborate more efficiently, but time remains a concern. The pace was set based on the amount of time team members had to dedicate to this project. The Consortium was not the first priority of everyone involved in the project. Good project management skills have helped to address this problem, as has the development of new collaboration tools.[44]

The Consortium struggled with data-sharing principles and standards. Negotiating a data-sharing plan with even a small number of companies was challenging and having a single legal liaison for each of the companies was found to be critical. Standard definitions were not all obvious. In some cases, people would fail to pass on crucial information before leaving for another position. However, in the end the project created a template for the Biomarkers Consortium for data-sharing plans, which should speed the work in subsequent projects. Also, FDA currently has an initiative to require uniform data submissions using standardized data fields, which would result in data that are much more amenable for sharing. Furthermore, health care reform is also expected to harmonize data practices, in part to reduce costs and improve care.[45]

The existing data had many limitations. The original studies were not designed to answer the research question investigated by the Consortium. The adiponectin data also had limitations because different companies used different assays to measure the protein, which required more work to ensure that the data could be combined reliably.

Broader issues also arose. The clarity of the research question is very important for defining the type of collaboration. The existence of a neutral convener—in this case the FNIH—was critical in gaining the trust of all the stakeholders involved in the project. Still, motivations were an issue. Depending on the question being asked, the openness of the contribution and of the output can change. In the case of the Biomarkers Consortium, the output is completely open, which is a good model for generating new knowledge. The nature of the collaboration also depends on whether it is developing standards and tools, aggregating data, creating new knowledge, or developing a product. Collaborations depend on trust and openness. Being clear about common goals, realizing the unique value each party brings to the effort, and striving for open inclusiveness can greatly improve collaborations.

The NEWMEDS Consortium

http://www.newmeds-europe.com

[44] Presentation by John Wagner, vice president for clinical pharmacology at Merck and Co., Inc., at the Workshop on Sharing Clinical Research Data.

[45] Ibid.

NEWMEDS, which is a project sponsored by the European Union, stands for Novel Methods for Development of Drugs in Depression and Schizophrenia. The NEWMEDS consortium was established to facilitate sharing of clinical trials data—in particular, coded participant-level data—from industry and academia to examine research questions in the precompetitive domain. The schizophrenia database, which includes data from AstraZeneca, Janssen, Eli Lilly, Lundbeck, and Pfizer, encompasses 64 industry-sponsored studies representing more than 25,000 patients, along with studies sponsored by the National Institute of Mental Health and the European Union. The depression database, with data from several of the same companies, includes 26 placebo-controlled, industry-sponsored studies covering more than 8,000 patients.[46]

Locating the data was a challenge. For example, companies are bought and sold, and products are exchanged among companies. Also, competing internal resources and priorities mean that data sharing is not necessarily the top priority. Compared with the YODA project's experience, de-identification was much less expensive and time consuming, requiring about 2 weeks of programming time. In the context of the amounts spent on clinical trials and the potential markets for new products, though, even rather expensive de-identification projects can be justified. The formulation of research questions and interpretation of data also need to be the result of active collaboration so that understandings are shared as well as data.[47]

A paradigm shift is occurring that redefines data sharing as an "ethical imperative." Studies should be given extra credit if they are willing to share data. This could be taken into account by Institutional Review Boards (IRBs), for instance, in judging the ethical validity of a study.

One Mind Initiative

http://1mind4research.org/programs

One Mind has numerous projects at various stages of development. The following are the flagship initiatives, which have been in development for over a year and are positioned for full operational launch in 2013.

The GEMINI Program – Knowledge Integration Network (KIN) for Traumatic Brain Injury (TBI) and Post-Traumatic Stress (PTS) – is a Public-Private Partnership combining best-in-class science, technology, and expertise. One Mind is catalyzing collaboration between international research centers of excellence, industry, and government to accelerate the translation of basic science into breakthrough diagnostics and improved treatments for TBI & PTS.

In 2013, One Mind will launch this multi-country, multi-site initiative to create a large-scale database of individuals with acute head trauma and/or psychological trauma with rigorous biomarker (e.g., genetics, imaging) and clinical measures to:

- Identify biological indicators of the causes and effects of diseases, or pathology.
- Investigate disease progression for earlier, more accurate diagnosis and treatment.
- Create new ways to share data between academia, industry, non-profits and government.

[46] Presentation made by Jonathan Rabinowitz, academic lead of NEWMEDS at Bar Ilan University, at the Workshop on Sharing Clinical Trial Data.

[47] Ibid.

- Empower patients to take a more active role in their own care, thereby contributing to the acceleration of research.

One Mind has joined the *Orion Bionetworks Alliance for Multiple Sclerosis*. Orion is a program of the Marin Community Fund, a non-profit 501(c)(3) corporation. Orion Bionetworks is a multidisciplinary computational modeling community that harnesses the power of data, technology, and collaboration to pave the way for Systems Medicine and transform the research and treatment of multiple sclerosis and other brain disorders.

They develop causal disease models with high predictive power that will drive progress towards better treatments and, ultimately, cures for multiple sclerosis and other devastating brain disorders. To advance the understanding of brain disorders, Orion Bionetworks is building powerful, data-driven disease models. To make real progress, they use systems biology to understand how numerous complex, nonlinear biological and patient factors interact to cause brain disorders.

A framework has been created to advance causal disease models using computational tools to integrate diverse biological and clinical data from registries and repositories. These disease models will be continually refined and improved through collection of new high-dimensional, quality data validated against real-world patient data. A unique cooperative alliance model enables sustainable, multidisciplinary collaboration.

No one entity has the full array of resources needed to advance the research and treatment of multiple sclerosis. In Orion BioNetworks, complementary stakeholders contribute patient data, expertise, computational models, IT capabilities or funds to support the coordinated development of next-generation models. All Alliance partners have access to data contributed to the Data Commons through a Data Exchange Portal with Analytics Capabilities.

PHASE I: The initial program integrates clinical, biomarker and imaging data from existing databases of over 7,000 patients into a first-generation causal MS disease model. Program partners: Accellerated Cure Project for MS; Brigham Women's Hospital, GNS Healthcare; MetaCell; PatientsLikeMe. Funding Partner: Janssen Pharmaceutical Research. This phase was successfully launched in January of 2013.

PHASE II: Use combined, prospectively collected imaging, biomarker and deep phenotypic data from an additional 2000-3000 patients.

PHASE III: Model 3.0 will be focused on identifying response biomarkers to MS treatments.

Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and UCSF Biobank

http://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx

The RPGEH, a scientific research program at Kaiser Permanente in California, is one of the largest research projects in the United States to examine the genetic and environmental factors that influence common diseases such as heart disease, cancer, diabetes, high blood pressure, Alzheimer's disease, asthma and many others.

The goal of the research program is to discover which genes and environmental factors—the air we breathe, the water we drink, as well as lifestyles and habits—are linked to specific diseases. This new knowledge has the potential to improve health and health care delivery by leading to new and improved diagnosis and treatment of disease and even prevention of some disease. One day your doctor may be even able to make a health care plan just for you based on your genetic profile and life experiences. This could include early testing for the diseases you might be likely to get, prescribing medications that will work best for you, and recommending lifestyle changes that will help keep you healthier.

Based on the over six million-member Kaiser Permanente Medical Care Plan of Northern California (KPNC) and Southern California (KPSC), the completed resource will link together comprehensive electronic medical records, data on relevant behavioral and environmental factors, and biobank data (genetic information from saliva and blood) from 500,000 consenting health plan members.

In addition to learning more about the genetic and environmental determinants of disease, Kaiser Permanente research scientists, working in collaboration with other scientists across the nation and around the world, hope to translate research findings into improvements in health and medical care. We also hope to develop a broader understanding of the ethical, legal, and social implications of using genetic information in health care.

The International Severe Adverse Events Consortium (iSAEC)

http://www.saeconsortium.org

The international Serious Adverse Event Consortium (iSAEC) is a nonprofit, bio-medical research organization founded in 2007. It is comprised of leading pharmaceutical companies, the Wellcome Trust, and academic institutions; with scientific and strategic input from the U.S. Food and Drug Administration (FDA) and other international regulatory bodies. The mission of the iSAEC is to identify DNA-variants useful in understanding the risk of drug-related serious adverse events (SAEs).

Patients respond differently to medicines and all medicines can have side effects in certain individuals. The iSAEC's work is based on the hypothesis these differences have (in part) a genetic basis, and its research studies examine the impact genetic variation on how individuals respond to a large variety of medicines. The iSAEC's initial studies have successfully identified genetic variants associated with drug-related liver toxicity (DILI) and Serious Skin Rashes (SSR). The majority of the iSAEC's genetic findings have been specific to a given drug versus across multiple drugs. However, a number of cross drug genetic alleles are starting to emerge that may provide important insights into the underlying biology/mechanism of drug induced SAEs (e.g. HLA*5701 or UGT1A1*28). iSAEC's findings clearly demonstrate an important role for the MHC genomic region (Chromosome 6), in the pathology of immunologically mediated SAEs such as DILI and SSR. They also emphasize the importance of immune regulation genes, in addition to a number of well characterized drug metabolism (ADME) genes.

In the iSAEC's second phase (2010-2015), novel, international clinical networks have been developed to deepen the understanding of the genetics of the following SAEs (across a diverse range of ethnic populations):

- Hepatotoxicity (DILI)
- Serious Skin Rash (DISI)

- Acute Hypersensitivity Syndrome (DRESS)
- Nephrotoxicity (DIRI)
- TdP/PQT effects (DITdP)
- Excessive Weight Gain (associated with Class 2 Anti-Psychotic Medications)
- IBD Therapy Related SAEs (4 different phenotypes), and
- Jaw Osteonecrosis (ONJ)

They will continue to apply state of the art genomic methods to fulfill these aims, including whole genome genotyping and next generation sequencing to explore the role "rare" genetic variants in drug induced SAEs.

In addition to supporting original genomic research on drug-related SAEs, the iSAEC is:

- Executing "open-use research practices and standards" in all its collaborations.
- Encouraging greater research efficiency and "speed to results" by pooling talent and resources, under a collaborative private sector leadership model, solely focused on research quality and productivity.
- Releasing all of its novel genetic discoveries/associations into the public domain uninhibited by intellectual property constraints.
- Enhancing the public's understanding of how the industry, academia and government are partnering to address drug-related adverse events.

The iSAEC assists in organizing and funding a variety of international research networks to aggregate case collection of appropriate scale and diversity (i.e. ethnicity and causal drug). They are conducting pilot initiatives that leverage electronic medical records and related databases to potentially identify individuals with relevant SAEs, in greater scale and number. These well-characterized clinical databases, from individuals who have experienced a SAE, are being compared with control cases to identify genetic variants that may be associated with the specific SAE. The identification of such genetic variants is believed to be essential to developing safer drugs, while also identifying patient populations for whom a medicine will have the greatest likelihood of providing medical benefits with the fewest risks.

Alzheimer's Disease Neuroimaging Initiative (ADNI)

http://adni-info.org

Alzheimer's disease (AD) affects almost 50% of those over the age of 85 and is the sixth leading cause of death in the US. Since 2005, the longitudinal Alzheimer's Disease Neuroimaging Initiative (ADNI) has been validating the use of biomarkers including blood tests, tests of cerebrospinal fluid, and MRI/ PET imaging for Alzheimer's disease (AD) clinical trials and diagnosis.

Now in its third phase (ADNI, ADNI GO and ADNI 2), ADNI 2 is studying the rate of change of cognition, function, brain structure, and biomarkers in 150 elderly controls, 450 subjects with mild cognitive impairment, 150 with mild to moderate AD and a new group of 100 people with significant, yet subtle, memory complaints, referred to as the significant memory concern cohort.

ADNI volunteers are the heart of the study and the most prevalent characteristic among them may be altruism. The participants make a multiyear commitment to a study that is providing the path toward treatment and prevention of AD while not offering any potential intervention. ADNI

is currently enrolling participants who have been diagnosed with mild-to-moderate AD and also seeking participants with subtle, yet distinct, memory concerns.

ADNI also maintains an unprecedented data access policy intended to encourage new investigation and to increase the pace of discovery in the race to prevent, treat, and one day cure AD. All data is made available without embargo. Armed with better knowledge of the first indications of AD from ADNI and other studies, researchers are beginning to test potential therapies at the earliest stages feasible when there may the greatest promise for slowing down progression of this devastating disease.

The web site – adni-info.org – is informational and is intended to provide an introduction of ADNI study basics. Scientists and researchers seeking access to ADNI data should visit UCLA's Laboratory of Neuroimaging ADNI database (ADNI LONI) (http://adni.loni.ucla.edu/). Later this year, the ADNI LONI database will be significantly enriched with the addition of whole genome sequences (WGS) for 800 ADNI participants.

ADNI researchers collect, validate and utilize data such as MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers as predictors for the disease. Data from the North American ADNI's study participants, including Alzheimer's disease patients, mild cognitive impairment subjects and elderly controls, are available from the site http://adni.loni.ucla.edu.

PatientsLikeMe

http://www.patientslikeme.com/about

PatientsLikeMe is a health information-sharing website for patients where they can form peer-to-peer relationships, establish profiles, provide and share health data, and make de-identified data available for research.[48]

A prominent mandate of the site is "give something, get something." If patients provide information for a research project, they should receive information in return that can help them make meaningful decisions.

Another motto is "patients first." In a data-sharing environment, the interests of the patients need to come first. "They have a lot more skin in this game than any of us in this room do ... They

[48] Sally Okun, health data integrity manager at PatientsLikeMe, described some of the lessons learned from the website during its 7 years of operation at the Workshop on Sharing Clinical Research Data.

have the expertise in managing [their conditions] that as clinicians and as researchers we could never have."

That observation leads to a third mandate: Listen well. Patients want to share their information. When patients were asked in a recent survey whether their health data should be used to help improve the care of future patients who have the same condition, 89 percent agreed (Alston, et al., 2012). Yet, when they were asked whether they thought their data were being shared, the majority said they either did not know or did not think so.

The data patients provide involve intimate parts of their daily lives. These patients are not simply human subjects; they are actually members of the research team. The biomedical research system has evolved to the point that all stakeholders can be involved in decisions. "Without patients, we would have no research. Let's start thinking about how we can best honor them, respect them, and allow them to develop the trust that they need to participate with us" (Alston, et al., 2012).

Distributed Systems for Clinical Research Information Sharing

This is an alternative to widespread data sharing. Sharing information derived from the data while minimizing the sharing of data themselves nullifies some of the barriers to data sharing.[49] Typical example is the *Query Health Initiative*, a system for sharing clinical information that has been promulgated by the Office of the National Coordinator for Health Information Technology. It uses the approach of sending the question to the data rather than bringing the data to the question. The question, in this case, is an executable program sent from the originator to the holder of data. The program then operates on a remote dataset and returns the answer to the sender.

An alternative approach based on the same idea, is to let a user log onto a remote system and do the analyses. The user needs to be able to access the system through a firewall, which many organizations are hesitant to permit. Other protections can be built into the system as well, such as a mechanism for determining whether the research has oversight by an IRB. A steering committee or IRB could be involved in reviewing and approving queries. Network management could provide for auditing, authentication, authorization, scheduling, permissions, and other functions. Local controls at the source of the data could monitor what kind of question is being asked, who is asking the question, and whether the question is worth answering.

A logical extension of such a system would be a multisite system in which research data from several different organizations are behind several different firewalls. A single question could be distributed to multiple sites and the responses compiled to produce an answer. Source data, such as information from electronic health records, could flow into research systems through firewalls. The result would be a system in which remote investigators can gain the information they need to answer a question while data are protected.[50]

Mini-Sentinel

[49] This approach was described by Richard Platt, professor and chair in the Department of Population Medicine, Harvard Medical School, and executive director of Harvard Pilgrim Health Care Institute at the Workshop on Sharing Clinical Research Data.

[50] Ibid.

http://mini-sentinel.org/about_us

Platt described a system developed by his group that implements this concept. The system, called *Mini-Sentinel*, is being used by FDA to do post-market medical product safety surveillance. It has a distributed database with data on more than 125 million people, 3 billion drug dispensing, and 2.4 billion unique patient encounters, including 40 million acute inpatient stays. Each of the 17 data partners involved in the project uses a common data format so that remote programs can operate on the data. Data checks ensure that the data are correct. Data partners have the option of stopping and reviewing the queries that arrive before the code is executed. They also can stop and inspect every result before it is returned to the coordinating center. The amount of patient-level data that is transferred is minimized, with most of the analysis of patient-level data done behind the firewall of the organization that has the data. "Our goal is not to never share data. Our goal is to share as little data as possible." The analysis dataset is usually a small fraction of all the data that exist, and the data can usually be de-identified.[51]

As an example of the kinds of projects that can be done using this system, Platt described a study looking at comparative risks of angioedema related to treatment with drugs targeting the renin-angiotensin-aldosterone system. The results of the study had not yet been released at the time of the workshop, but their experience has shown that data from millions of people could be accessed to do the study without sharing any patient-level data. Yet, from the perspective of the investigators, "essentially everything that was interesting in those datasets that could answer this question was accessible and was used to address the questions of interest."

Using such a system, it would be possible to address a large fraction of the questions thought to require data sharing by instead sharing programs among organizations that are prepared to collaborate on distributed analyses. Organizations also could participate in multiple networks, further expanding the uses of the data they hold. At the same time, every network could control its own access and governance.

Today, only FDA can submit questions to Mini-Sentinel, but FDA believes it should be a national resource and is working on ways to make it accessible to others. Toward that end, the week before the workshop, the NIH announced the creation of the Health Care Systems Research Collaborative, which will develop a distributed research network with the capability of communicating with the Mini-Sentinel distributed dataset. Such systems, by sharing information rather than data, could make progress faster than waiting for all the issues surrounding data sharing to be resolved.

tranSMART

http://www.transmartproject.org

This project is an initiative of Johnson & Johnson. The company intended to bring together data and informatics across their immunology, oncology, and biotechnology franchises, which originally had been different companies with many different clinical trials and standards. Rather than reinventing the wheel, they built their system off a data warehousing tool called i2b2 that had been developed by researchers at Harvard for data from electronic health records. They made it open source and ran it through Amazon's cloud computing service.[52]

[51] Ibid.

[52] Perakslis at the Workshop on Sharing Clinical Trial Data.

tranSMART is a knowledge management platform that enables scientists to develop and refine research hypotheses by investigating correlations between genetic and phenotypic data, and assessing their analytical results in the context of published literature and other work.

The integration, normalization, and alignment of data in tranSMART permits users to explore data very efficiently to formulate new research strategies. Some of tranSMART's specific applications include: (1) Revalidating previous hypotheses; (2) Testing and refining novel hypotheses; (3) Conducting cross-study meta-analysis; (4) Searching across multiple data sources to find associations of concepts, such as a gene's involvement in biological processes or experimental results; (5) Comparing biological processes and pathways among multiple data sets from related diseases or even across multiple therapeutic areas

The tranSMART Data Repository combines a data warehouse with access to federated sources of open and commercial databases. tranSMART accommodates:

- Phenotypic data, such as demographics, clinical observations, clinical trial outcomes, and adverse events
- High content biomarker data, such as gene expression, genotyping, pharmacokinetic and pharmaco-dynamics markers, metabolomics data, and proteomics data
- Unstructured text-data, such as published journal articles, conference abstracts and proceedings, and internal studies and white papers
- Reference data from sources such as MeSH, UMLS, Entrez, GeneGo, Ingenuity, etc.
- Metadata providing context about datasets, allowing users to assess the relevance of results delivered by tranSMART

Data in tranSMART is aligned to allow identification and analysis of associations between phenotypic and biomarker data, and it is normalized to conform with CDISC and other standards to facilitate search and analysis across different data sources. tranSMART also enables investigators to search published literature and other text sources to evaluate their analysis in the context of the broader universe of reported research.

External data can also be integrated into the tranSMART data repository, either from open data projects like GEO, EBI Array Express, GCOD, or GO, or from commercially available data sources. Making data accessible in tranSMART enables organizations to leverage investments in manual curation, development costs of automated ETL tools, or commercial subscription fees across multiple research groups.

tranSMART's Dataset Explorer provides flexible, powerful search and analysis capabilities. The core of the Dataset Explorer integrates and extends the open source i2b2 application, Lucene text indexing, and GenePattern analytical tools.

Connections to other open source and commercial analytical tools such as Galaxy, Integrative Genomics Viewer, Plink, Pathway Studio, GeneGo, Spotfire, R, and SAS can be established to expand tranSMART's capabilities.

tranSMART's design allows organizations flexibility in selecting analytical tools accessible through the Dataset Explorer, and provides file export capabilities to enable researchers to use tools not accessible in the tranSMART portal.

The tranSMART system also incorporates role-based security protections to enable use of the platform across large organizations. User authentication for tranSMART can be integrated into an organization's existing infrastructure to simplify user management.

The tranSMART security model allows an organization to control data access and use in accordance with internal policies governing the use of research data, as well as HIPAA, IRB, FDA, EMEA, and other regulatory requirements.

A growing number of biopharmaceutical companies, academic medical centers, and other research-oriented organizations are joining the tranSMART community, affirming the 2010 Bio-IT World Best Practices and CIO 100 awards granted for the platform's innovation and potential to advance translational research.

Several lessons emerged from the experience. First, to use the standards that are available because "patients are waiting." At some point, human curators are going to be necessary to align the data and insert it into a database, but to get the project moving forward, one should start with what already works. Second, an important goal for a project such as this one is to rule out options quickly. Clinical trials should not waste patients' time on drugs that are not going to work.

It can be concluded from the experience that light and agile data "marts" are preferable or databases generated to answer specific questions or test hypotheses, over large data warehouses. It is better to aggregate the source around the question quickly and effectively. That way, as technologies, standards, and definitions change, tools are flexible and can change accordingly.

CAMD (Coalition Against Major Diseases) and C-Path Alzheimer's Database

http://www.c-path.org/News/CDISCTAStds%20PR-24June2012.pdf

This is an example of a data-sharing approach that has benefited from the use of standards. Without standards, integrating datasets and pooling data is difficult. The Critical Path Institute acts as a trusted third party that works with partners in FDA, industry, and academia to develop consensus measurement, method, and data standards. The standards are then submitted to FDA for qualification. After qualification is achieved, the standards enter the public domain and can be used by everyone. C-Path "convenes consortiums to bring together the best science and in this fashion create shared risk and shared cost for the creation of these standards.

One of the six such global consortiums organized by C-Path is the *Coalition Against Major Diseases* (CAMD). This consortium focuses on diseases of the brain and peripheral nervous system. The coalition is seeking to advance drug development tools and standards as a means of addressing the challenge of the unsustainable time and cost required to get a new drug to market. In particular, the focus of its efforts is process improvement to advance effective treatments for Alzheimer's and Parkinson's diseases. CAMD is working to qualify biomarkers as

drug development tools and has also been developing standards to create integrated databases drawn from clinical trials. These databases have been used to model clinical trials to optimize trial design.

Nine member companies agreed to share placebo control data from 22 clinical trials on Alzheimer's disease, but the data were not in a common format and needed to be combined in a consistent manner, Compton explained. All data were remapped to the CDISC standards and pooled. The resulting database was used to develop a new computerized clinical trial simulation and modeling tool. To get there, however, the contributing companies had to go through a corporate approval process to share and de-identify the data, after which C-Path did further de-identification to ensure compliance with Health Insurance Portability and Accountability Act requirements.

The modeling tool allowed for accurate quantitative predictions of defined patient populations. By merging data from diverse sources, 65-year-old males who looked alike in the databases could be divided into three classes with different trajectories of disease. "Seeing this kind of distinction emerge from the modeling tool would allow you to design a trial much more wisely. "It would inform patient selection, study size, study duration, study feasibility, and even study costs." [53]

Several key insights have been gained from the project. First, legacy data conversion is resource dependent, but worthwhile for specific projects. In this case, de-identifying data and converting it to a standard format took 9 months, but generated a database with 6,100 Alzheimer's disease patients. To get the value back from the conversion process, it is important to assess upfront that the database will be useful for achieving specific objectives, like qualifying a new tool. If it will be, selectivity is beneficial. It's better to convert the data one needs, [but] maybe not everything. Once data are converted to a common standard and aggregated, the addition of standardized data from other sources, whether prospective or retrospective, becomes simplified and expands the power and utility of a standardized data resource. The database continues to grow over time and in power.

Based on the success with Alzheimer's, the approach is now being applied to other research projects, including the development of new tools for Parkinson's disease, polycystic kidney disease, and tuberculosis. This approach could cut drug development times "by 4 to 5 years." Such tools also have applications to post-approval safety monitoring and data gathering.[54]

Clinical Trials Network (CTN) Data Share

http://www.ctndatashare.org/index

To date, the efficacy of new treatments for drug addiction has been demonstrated primarily in specialized research settings, with somewhat restricted patient populations. To address this problem, the U.S. National Institute on Drug Abuse (NIDA) established the National Drug Abuse Treatment Clinical Trials Network (CTN).

[53] Conclusions made by Carolyn Compton, president and chief executive officer of the Critical Path Institute (C-Path), at the Workshop on Sharing Clinical Trial Data.
[54] Ibid.

The Clinical Trials Network (CTN) Data Share web site is an electronic environment that allows data from completed clinical trials to be distributed to investigators in order to promote new research, encourage further analyses, and disseminate information to the community. Secondary analyses produced from data sharing multiply the scientific contribution of the original research. This site allows researchers to download de-identified data from completed CTN studies to conduct analyses that improve the quality of drug abuse treatment.

Effective data sharing includes communicating to the research community that data are available, providing sufficient descriptive information about the data, enforcing compliance to standard semantics and structure, rendering the data in a usable format, and making data accessible.

A primary concern in sharing data is the protection of human subjects. The rights and privacy of people who participate in NIH-sponsored research must be protected at all times. Thus, data on this site have been completely de-identified to prevent linkages to individual research participants. This includes removal of all Personal Health Information (PHI) and indirect identifiers that are not listed as PHI but could lead to "deductive disclosure" such as comment fields and site numbers. Study-specific de-identification methods are documented with each protocol.

Data are available in either a Clinical Data Interchange Standards Consortium (CDISC) format or a Case Report Form (CRF) format. For some studies, both formats are available. For the CDISC format, prior to de-identifying the data, all data files are converted from their native format to a modified Study Data Tabulation Model (SDTM) standard format. The SDTM is a content standard sponsored by CDISC. This universal data format will be applied to each completed CTN trial's data. This will facilitate the pooling of shared data across completed studies, as the variable names are consistent across studies. For the CRF format, separate data files are created for each CRF collected on the study. This will facilitate those researchers interested in looking at all data from a single CRF in one data file, as the data files match the CRF exactly.

De-identified data are available for download in two formats: SAS (transport files) and ASCII (CSV files). Documentation regarding the data and corresponding study that generated the data are also available under each completed protocol page. This includes the annotated case report form (CRF); a define.xml file outlining the structure, variables and contents of each dataset; and SDTM mapping for the CDISC data and de-identification rules.

Prior to downloading any study data, the user will be prompted to complete a registration agreement for data use. Users will have to register a name, position, affiliation, valid email address, and country of origin in order to download data to accept their responsibility for using data in accordance with the CTN Data Share Agreement.

Data for completed protocols will be available to the public approximately 18 months after completion of the study or after the acceptance of publication of the primary manuscript, whichever occurs first.

The "ePlacebo" Database

http://www.genome.gov/19518664

Michael Cantor, senior director of clinical informatics and innovation at Pfizer Inc., described during the workshop on sharing clinical trial data an ongoing data-sharing project being undertaken by Pfizer as part of its "Data Without Borders" initiative. The project, called ePlacebo, pools data from placebo and control arms across multiple clinical trials in a variety of therapeutic areas. The result is a large comparison group that can be used to evaluate events that might not be seen in a single trial, study placebo effects, and possibly reduce the size of placebo arms needed in future clinical trials. So far, data from about 20,000 patients have been compiled from hundreds of trials, and Pfizer is hoping to expand the utility of this data source by soliciting participation from other organizations.

The goal for ePlacebo is to provide a resource that is inclusive, rests on standards, and spans disease areas. The intent is to set it up as a self-service dataset that could be used for any legitimate research purpose. However, consistent data standards have only been implemented at Pfizer within the past decade and as a result, only relatively recent studies were used for ePlacebo because of the difficulties combining data from trials that did not use standards or implemented them in different ways.

Innovative Medicines Initiative (IMI) European Medical Information Framework (EMIF)

http://www.imi.europa.eu/content/home

The Innovative Medicines Initiative (IMI) is Europe's largest public-private initiative aiming to speed up the development of better and safer medicines for patients. IMI supports collaborative research projects and builds networks of industrial and academic experts in order to boost pharmaceutical innovation in Europe. IMI is a joint undertaking between the European Union and the pharmaceutical industry association EFPIA.

Stanford Microarray Database

http://smd.princeton.edu/

The Stanford Microarray Database (SMD) stores raw and normalized data from microarray experiments, and provides web interfaces for researchers to retrieve, analyze and visualize their data. The two immediate goals for SMD are to serve as a storage site for microarray data from ongoing research at Stanford University, and to facilitate the public dissemination of that data once published, or released by the researcher. Of paramount importance is the connection of microarray data with the biological data that pertains to the DNA deposited on the microarray (genes, clones etc.). SMD makes use of many public resources to connect expression information to the relevant biology, including SGD (Ball, et al., 2000), YPD and WormPD (Costanzo, M.C., et al., 2000), Unigene (Wheeler, et al., 2000), dbEST (Boguski, Lowe, & Tolstoshev, 1993), and SWISS-PROT (Bairoch & Apweiler, 2000).

ArrayExpress – Functional Genomics Data

http://www.ebi.ac.uk/arrayexpress/

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MINSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set

http://discover.nci.nih.gov/cellminer/

High-throughput and high-content databases are increasingly important resources in molecular medicine, systems biology, and pharmacology. However, the information usually resides in unwieldy databases, limiting ready data analysis and integration. One resource that offers substantial potential for improvement in this regard is the NCI-60 cell line database compiled by the U.S. National Cancer Institute, which has been extensively characterized across numerous genomic and pharmacologic response platforms. The Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology (LMP), Center for Cancer Research (CCR), National Cancer Institute (NCI), has developed a CellMiner (http://discover.nci.nih.gov/cellminer/) – web application designed to improve the use of this extensive database. CellMiner tools allowed rapid data retrieval of transcripts for 22,379 genes and 360 microRNAs along with activity reports for 20,503 chemical compounds including 102 drugs approved by the U.S. Food and Drug Administration. Converting these differential levels into quantitative patterns across the NCI-60 clarified data organization and cross-comparisons using a novel pattern match tool. Data queries for potential relationships among parameters can be conducted in an iterative manner specific to user interests and expertise. Examples of the in silico discovery process afforded by CellMiner were provided for multidrug resistance analyses and doxorubicin activity; identification of colon-specific genes, microRNAs, and drugs; microRNAs related to the miR-17-92 cluster; and drug identification patterns matched to erlotinib, gefitinib, afatinib, and lapatinib. CellMiner greatly broadens applications of the extensive NCI-60 database for discovery by creating web-based processes that are rapid, flexible, and readily applied by users without bioinformatics expertise (Reinhold, et al., 2012).

Structural Genomics Consortium (SGC)

http://www.thesgc.org

SGC is a public-private partnership that supports the discovery of new medicines through open access research. The core mandate of the SGC is to determine 3D structures on a large scale and cost-effectively – targeting human proteins of biomedical importance and proteins from human parasites that represent potential drug targets. The SGC is now responsible for >10% of all novel human protein structures deposited into the Protein Data Bank each year; to May 2013, the SGC has released almost 1500 structures of proteins with implications to the development of new therapies for cancer, diabetes, obesity, and psychiatric disorders.

In this effort, the SGC has gained a reputation for the quality and the reproducibility of its research output, and for meeting its milestones on time and on budget. In its current configuration, the SGC includes active research facilities at the Universities of Toronto and Oxford. Current funders of the SGC include AbbVie, Boehringer Ingelheim, the Canada

Foundation for Innovation, the Canadian Institutes for Health Research, Genome Canada, GlaxoSmithKline, Janssen, Lilly Canada, the Novartis Research Foundation, the Ontario Ministry of Economic Development and Innovation, Pfizer, Takeda, and the Wellcome Trust. Recently, these organizations together have committed greater than US$72 Million to the consortium to sustain operation for the next 4 years.

Since its inception, the SGC has been engaged in pre-competitive research to facilitate the discovery of new medicines. A vital part of the SGC's missons is to generates reagents and knowledge related to human proteins and proteins from human parasites.

More importantly, the SGC has adopted an Open Access policy as it believes that all of its output (knowledge, data and reagents) will have maximal benefit if released into the public domain without restriction on use. This enabled a very inclusive model for collaborations worldwide, ultimately speeding up early stage drug discovery. Open Access also allows us to reach beyond the scientific community, engaging clinicians and patient groups for instance.

The database of Genotypes and Phenotypes (dbGaP)

http://www.ncbi.nlm.nih.gov/gap

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The advent of high-throughput, cost-effective methods for genotyping and sequencing has provided powerful tools that allow for the generation of the massive amount of genotypic data required to make these analyses possible.

dbGaP provides two levels of access—open and controlled—in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Summaries of studies and the contents of measured variables as well as original study document text are generally available to the public, while access to individual-level data including phenotypic data tables and genotypes require varying levels of authorization.

The data in dbGaP will be pre-competitive, and will not be protected by intellectual property patents. Investigators who agree to the terms of dbGaP data use may not restrict other investigators' use of primary dbGaP data by filing intellectual property patents on it. However, the use of primary data from dbGaP to develop commercial products and tests to meet public health needs is encouraged.

Open-access data can be browsed online or downloaded from dbGaP without prior permission or authorization. These data will include, but may not be limited to, the following: Studies, Study Documents, Phenotypic Variables, Genotype-Phenotype Analyses.

Sage Bionetworks

http://sagebase.org/info/index.php

Sage Bionetworks' moto is "We are redefining biomedical research through open systems, incentives, and norms."

They work to "redefine how complex biological data is gathered, shared and used." They challenge the traditional roles of individuals and groups, patients and researchers. Sage Bionetworks' work includes the building of platforms and services and undertaking research developing predictors relating to health. Arising in 2009 from Rosetta Inpharmatics (Merck Inc), Sage Bionetworks is a non-profit research organization based in Seattle, US and collaborates with a worldwide network. Sage Bionetworks is expanding its interdisciplinary team to accelerate the development of network models of biology across a variety of disease areas and species.

Sage Bionetworks create technology platforms that facilitate collaboration on data, governance platforms that enable data sharing and reuse, run challenges to solve complex biomedical problems, and perform own cutting-edge computational biology and research.

Sage Bionetworks works on the tools and platforms required to gather, share and use data in novel ways. These are targeted both at the research community, as well as organizations and individuals who are involved in providing data and being involved in the research process. They range from the technology platforms, Synapse and BRIDGE, through novel methods of addressing governance issues and Portable Legal Consent, to the ability to run Challenges to solve data-driven questions.

Synapse is a platform to support open, collaborative data analysis for clear, reproducible science. Synapse is currently targeting scientists working with clinical and genomics data. Synapse is currently released, with ongoing modifications and improvements planned in coming months. It currently hosts over 10,000 data sets available for use. One can set up a free account and start collaborating, or can download the source code at Github, view company's developer documentation, and get started. The software is available in Github, and the non-software creative works are licensed under the Creative Commons Attribution 3.0 Unported license except for legacy publications in closed journals.

The Synapse platform is composed of a set of shared web services that support a website designed to facilitate collaboration among scientific teams, and integrations with analysis tools and programming environments.

Synapse goes beyond being a data repository and creates a computational platform for real-time research collaboration. The company has begun to see evidence that a shared environment can help scientific teams find and correct analysis errors more quickly, get more diverse teams working on complex problems, and integrate disparate data to answer scientific questions.

The organizing construct of the portal revolves around allowing users to define their own online project spaces to which they can post content (data, code, analysis history and results) and document their work online immediately upon production. Project owners are able to control access to their own projects: users can share projects initially with only defined collaborators or make them publicly available.

Synapse is now actively rolling out to support computational research. It is the official resource for hosting analysis-ready data for use in the TCGA pan cancer consortium analysis working group, and will provide the versioned, citable datasets referenced in the TCGA Pan Cancer publications from the Nature Publishing Group to appear as a collection later this year. The Mt. Sinai School of Medicine has also chosen Synapse to support groups working in Alzheimer's disease and diabetes.

Bridge is an open system which is built around a Commons marketplace where patient citizens can donate their data and track its use and engage with researchers to best define and understand the questions important to their community.

At its heart it is about bringing the people with the best knowledge of their condition together with the people interested in researching their disease or condition in an open user-driven community.

This can have many benefits for both the researchers and the individual or disease communities including gathering rich data from a wide participant base, sharing of data and research with a wider audience, and gaining the time and efficiency benefits from a crowdsourcing approach to the research.

Bridge is really defined by the projects currently ongoing or planned. The initial pilot projects are working with the Fanconi Anaemia and Breast Cancer communities.

Parkinson's Progression Markers Initiative (PPMI)

http://ppmi-info.org

The mission of PPMI is to identify one or more biomarkers of Parkinson's disease (PD) progression, a critical next step in the development of new and better treatments for PD. PPMI will establish a comprehensive, standardized, longitudinal PD data and biological sample repository that will be made available to the research community.

PPMI is an observational, multi-center study that will collect clinical and imaging data and biologic samples from various cohorts, that can be used by scientists to establish markers that can be used by scientists to establish markers of disease progression in PD.

Specific aims to accomplish this objective include:

- Develop a comprehensive and uniformly acquired clinical/imaging dataset with correlated biological samples that can be used in biomarker verification studies
- Establish standardized protocols for acquisition, transfer and analysis of clinical and imaging data and biological samples that can be used by the research community
- Investigate existing biomarkers and identify new clinical, imaging and biological markers to determine interval changes in these markers

Biosense (CDC)

http://www.cdc.gov/biosense

The BioSense program tracks health problems in the United States as they evolve. It provides public health officials with the data, information, and tools needed to better prepare for and coordinate responses to safeguard and improve the health of Americans.

BioSense is a unique public health tool that provides a picture of what is happening right now with any health condition, anywhere and everywhere in the country. BioSense pulls together information on emergency department visits and hospitalizations from multiple sources, including the Department of Veterans Affairs (VA), the Department of Defense (DoD), and civilian hospitals around the country. The BioSense program works with state or local health departments that have agreed to share data from their own emergency department monitoring systems to collect data from civilian hospitals.

Analysis of data through BioSense provides insight into the health of communities across the country. Such data are vital to guide decision making and actions by public health agencies at local, regional, and national levels.

The National Cancer Informatics Program (NCIP) Open-Development Initiative

http://cbiit.nci.nih.gov/ncip

The NCIP is committed to improving biomedical informatics through broad community participation. To this end, NCIP has deposited a large volume of open-source code for biomedical informatics software applications to GitHub Site Exit Disclaimer, a widely used code-hosting site. GitHub allows community developers to easily access and modify the code for applications of interest, and to contribute these modifications back to the primary repository for broader community dissemination. Most of these projects are available under the standard BSD 3-clause license to encourage both open-source and commercially viable derivatives of these projects.

The NCI Biomedical Informatics Blog provides a forum for the exploration of ideas and knowledge sharing across the NCI and throughout the wider biomedical-informatics and cancer-research community.

Topics recently under discussion include potential NCI Cancer Knowledge Clouds; NCIP support for open, community-driven software development; and intramural collaborations being pursued by NCIP staff. For a more complete description, scroll through the running list of topics contained in the left navigation bar.

The NCI Wiki hosts individual wikis that document the development of applications, activities of working groups, collaborations, informatics projects, and work in specific domains. Visit the wiki dashboard and select the New Account link to participate. For more information, contact Application Support.

Cancer Commons

http://www.cancercommons.org/about

Cancer Commons is a nonprofit, open science initiative linking cancer patients, physicians, and scientists in Rapid Learning Communities. Its mission is to ensure that patients are treated in

accord with the latest knowledge on targeted therapies and immunotherapies and to continually update that knowledge based on each patient's response.

Advisory boards, comprising leading physicians and scientists in each cancer, curate the molecular model that identifies the most relevant tests, treatments, and trials for each molecular subtype of that cancer, based on the best current knowledge.

Patients and physicians access knowledge derived from the model, as soon as it's deemed actionable, through alerts, notifications, news, and applications that inform testing and treatment decisions.

Physicians, researchers, and patients report clinical observations and outcomes and new data are analyzed in our Rapid Learning Communities by collaborative, inter-institutional teams to validate, refute, and amplify the current model's knowledge and hypotheses. The advisory boards reviews these results and update the framework in real time.

5 OpenScienceLink Key Performance Indicators (KPIs)

5.1 KPIs purpose

The Key Performance Indicators (KPIs) presented in this document focus on the monitoring of the progress of the **project's developments**, **pilot operations** and **sustainability efforts** and are differentiated from the KPIs of the OpenScienceLink services and tools which will be covered in WP8. Hence, their primary purpose is to comprise the means for monitoring the project's work towards the realisation of its objectives; both the general and the specific ones. The following table summarises these objectives (as analysed in the OpenScienceLink Description of Work).

Objective ID	Objective Description	Related WPs
General Objectives		
General Objective 1 (GO1)	To **design**, **implement**, **pilot**, **evaluate**, **expand** and **sustain** an ecosystem enabling multiple stakeholders to **access**, **search**, **process**, **link**, **analyze** and **evaluate openly accessible scientific information** as well as **collaborate** for the **generation**, **review** and **evaluation** of **related scientific information**. Towards this direction, leading edge ICT technologies in the areas of semantic search, data mining and Web2.0/Web3.0 social networking will be exploited through the integration of existing platforms and tools.	WP2, 3, 4, 5, 6, 7, 8
General Objective 2 (GO2)	To **introduce new added-value processes for creating and managing data journals**, **identifying and using research trends**, **creating networks for researchers clustering** and **collaboration**, **reviewing scientific information**, **designing and calculating metrics of scientific performanc**e.	WP2, 3, 4, 5, 6, 7, 8
General Objective 3 (GO3)	To **document** and **promote** a set of **best practices** and **blueprints** for the **optimal exploitation and use of openly accessible scientific information**. These best practices will include **novel business models** based on **open access to scientific information**, while also boosting the development of relevant **policies** that could encourage the creation, publishing and management of associated **open data**. Also, the best practices will cover the ever important **legal aspects**, including IPR management associated with openly accessible scientific information.	WP8
Specific Objectives		
Specific Objective 1 (SO1)	To **elicit**, **understand** and **analyze requirements** associated with the creation, use, management and linking of openly accessible scientific information, *including the perspectives of all relevant stakeholders* including publishers, researchers, scholars, research organizations, universities, research sponsors, funding authorities and policy makers.	WP2, 3
Specific Objective 2 (SO2)	To **integrate** a novel **ICT platform** based on readily available research results from **background R&D projects** (namely FP7 projects PONTE and SocIoS), as well as the GoPubMed semantic search engine. This ICT platform will provide the **advanced semantic search**, **data mining** and	WP5,7,8

Objective ID	Objective Description	Related WPs
	social networking functionalities required to support the OpenScienceLink pilot services. The platform will be integrated, **tested** and **piloted** on the basis of a wide range of **open data repositories** in the area of **medical, clinical and biological research**. Also, the platform will incorporate a **framework** (including a range of ICT-based tools) for the management of the ever important **legal and IPR aspects**.	
Specific Objective 3 (SO3)	To **specify**, **pilot**, **validate** and **evaluate** processes for **developing**, **searching**, **classifying** and **linking** (openly accessible) **data journals**, along with related process for using data journals for research benchmarking purposes.	WP3, 4, 5, 7, 8
Specific Objective 4 (SO4)	To **specify**, **pilot**, **validate** and **evaluate** added-value processes for **reviewing research articles**.	WP3, 4, 5, 7, 8
Specific Objective 5 (SO5)	To **specify**, **pilot**, **validate** and **evaluate** added-value processes for **identifying research trends** and managing **metrics of scientific performance**.	WP3, 4, 5, 7, 8
Specific Objective 6 (SO6)	To **establish collaboration networks** between researchers and to **validate** and **evaluate** the **added-value of such networks** for **researchers' collaboration**, **exchange of experiences** and **knowledge sharing**.	WP5, 7, 8
Specific Objective 7 (SO7)	To **sustain** the OpenScienceLink platform and associated ecosystem, while at the same time **plan its gradual expansion** through its enhancement with additional information and researchers from other research areas.	WP9
Specific Objective 8 (SO8)	To **elicit, document and promote best practices** and **blueprints** for **using/exploiting openly accessible scientific information**. These best practices will be **actively disseminated** to **decision makers and policy makers** in order contribute to the shaping of policies that will encourage and promote open access to scientific information. Special emphasis will be paid on the specification of a **legal framework** (including IPR aspects) for regulating **access** and **use** of **openly available scientific information across a variety of scenarios**, starting from the five pilot scenarios/services to be piloted in the project.	WP8

Table 32: OpenScienceLink Objectives

5.2 Project KPIs

This section presents the KPIs which are used for the monitoring and evaluation of the project's activities and are associated with the planning and implementation of the OpenScienceLink platform, as well as their sustainability and wider use. The list of potential KPIs presented here is structured along four main perspectives of a possible Balanced Scorecard for the OpenScienceLink project, with a view on sustainability, i.e. by addressing also follow-up efforts.

The four perspectives of the Balanced Scorecard are:

- **Mission perspective**: "How we create value for our stakeholders in the holistic approach to the publication, sharing, linking, review, and evaluation of research results?"
- **Resource perspective**: "How do we add value to stakeholders while controlling costs?"
- **Internal Business processes**: "How good we are in performing the core business processes?"
- **Learning & Growth perspective**: "How do we sustain the capacity to grow and innovate, meeting new and emerging demands?"

Taking into consideration these perspectives, the following table presents a set of quantitative KPIs related to the above along with the expected result each one of them serves, the method which will be used for its measurement, the project objectives it targets at and its (**cumulative**) target values on a yearly basis, including one year period after the project's completion.

Ind. ID	KPI Name	Objective/ Expected Result	Method of Measurement	Related Project Objective(s)	Expected Progress			
					Year1	*Year2*	*Year3*	*Year4*[55]
Mission perspective Reach out: *Number of Pilot Users*								
KPI1	**Number of Researchers / Scholars registered** in the OpenScienceLink Pilot Operations	**Credible Large Scale Platform and Services Validation / User Participation**	Track unique users in the platform	GO1, SO1, SO2, SO6, SO7	0	>=400	>=1100	>=2000
KPI2	**Number of Researchers /Scholars using** the system **at least once per month within a 3 months period**	**Credible Large Scale Platform and Services Validation / Active User Participation**	Track unique users in the platform using the system at least once per month in 3 months period	GO1, SO1, SO2, SO6, SO7	0	>= 150	>= 600	>= 1000
KPI3	**Number of Universities/ Academic Institutions** (beyond	**Credible Large Scale Platform and Services Validation /**	Track universities/ institutions in the	GO1, SO1, SO2, SO5,	>=10	>=20	>=30	>=80

[55] One year following the end of the project

Ind. ID	KPI Name	Objective/ Expected Result	Method of Measurement	Related Project Objective(s)	Expected Progress			
					Year1	*Year2*	*Year3*	*Year4*[55]
	the Consortium) **involved** in the OpenScienceLink Pilot Operations	**Stakeholders Participation**	platform	SO6, SO7				
KPI4	**Number of Research Sponsors and/or Funding Authorities** (beyond the Consortium) **involved** in the OpenScienceLink Pilot Operations	**Credible Large Scale Platform and Services Validation / Stakeholders Participation**	Track Sponsors engaging with the project and its results	GO1, SO5, SO7	0	>=2	>=5	>=10
Reach out: *Geographic coverage*								
KPI5	**Number of Countries** from which the OpenScienceLink Platform has been **used at least once**	**Credible Large Scale Platform and Services Validation / International Participation**	Track countries from which users of the platform come	GO1, SO1, SO2, SO6, SO7	0	>=10	>=30	>=40
KPI6	**Number of Countries** with **active** OpenScienceLink users (i.e., **using** the system **at least once per month within a 3 months period**)	**Credible Large Scale Platform and Services Validation / Active International Participation**	Track countries from which active users of the platform come	GO1, SO1, SO2, SO6, SO7	0	>=9	>=25	>=35
Reach out: *Domain coverage*								
KPI7	**Number of Biomedical and Clinical Research areas** (such as cardiology, pharmacology, etc) in which **researchers registered** to the OpenScienceLink platform belong to	**Credible Large Scale Platform and Services Validation / Research Domain Involvement**	Track distinctive biomedical and clinical research domains in which registered users of the platform are involved	GO1, SO1, SO2, SO6, SO7	0	>=8	>=10	>=15
KPI8	**Number of Biomedical and Clinical Research areas** (such	**Credible Large Scale Platform and Services Validation /**	Track distinctive biomedical and	GO1, SO1, SO2, SO6, SO7	0	>=5	>=8	>=12

Ind. ID	KPI Name	Objective/ Expected Result	Method of Measurement	Related Project Objective(s)	Expected Progress			
					Year1	*Year2*	*Year3*	*Year4*[55]
	as cardiology, pharmacology, etc) in which **researchers** having been **actively using** the OpenScienceLink platform	**Research Domain Active Involvement**	clinical research domains in which active users of the platform are involved (i.e., using the system at least once per month within a 3 months period) belong to					
Resource perspective *Attraction of additional resources for furthering the OpenScienceLink achievements in the holistic approach to the publication, sharing, linking, review, and evaluation of research results*								
KPI16	**Number of funded projects** *(beyond OpenScienceLink)*		Track projects in which the OpenScienceLink services, models and tools are used and further developed	GO1, SO7	9	9	>=11	>=12
KPI17	**Number of Stakeholders** involved in the partnership		Track Stakeholders introduced in the partnership	GO1, SO7	9	9	9	>=10
KPI18	**Number of Stakeholders** having formally expressed commercial interest		Track Stakeholders expressing commercial interest (through letters of interest, establishment of teleconferences, etc)	GO1, SO7	0	0	>=3	>=5

Ind. ID	KPI Name	Objective/ Expected Result	Method of Measurement	Related Project Objective(s)	Expected Progress			
					Year1	*Year2*	*Year3*	*Year4*[55]
Project Visibility and Interested Parties' Awareness								
KPI19	**Number of Journal Publications**	**Broad dissemination of the project results in scientific journals**	Track OpenScienceLink-related scientific papers published at journals by Consortium members	GO1, SO7	>=1	>=4	>=8	>=9
KPI20	**Number of Publications in Blogs and Web Sites**	**Broad dissemination of the project results in the web**	Track posts at blogs and web sites about OpenScienceLink	GO1, SO7	>=1	>=3	>=5	>=10
KPI21	**Number of Conference, Workshop and Exhibitions** Publications and Presentations	**Broad dissemination of the project results in conferences, workshops and exhibitions**	Track OpenScienceLink-related papers, presentations and presence at conferences, workshops and exhibitions	GO1, SO7	>=3	>=7	>=13	>=16
KPI22	**Joint workshops with other projects** or related national initiatives	**Strengthened Collaboration with relevant projects and initiatives**	Track workshops held with other projects or related initiatives	GO1, SO7	0	>=2	>=5	>=5
KPI23	**Number of OpenScienceLink Unique Website Visitors**	**Broad dissemination of the project results through the website**	Track OpenScienceLink website visitors	GO1, SO7	>=100	>=200	>=400	>=500
KPI24	**Number of external websites** referring to OpenScienceLink website	**Project results Publicity across the Internet**	Track references to OpenScienceLink website from	GO1, SO7	>=10	>=15	>=20	>=25

Ind. ID	KPI Name	Objective/ Expected Result	Method of Measurement	Related Project Objective(s)	Expected Progress			
					Year1	*Year2*	*Year3*	*Year4*[55]
			external ones					
KPI25	**Number of Press Releases** published	**Broad dissemination of the project results in press**	Track OpenScienceLink-related press releases published by Consortium members	GO1, SO7	>=2	>=5	>=10	>=15
KPI26	**Number of Participations in relevant** (open access) **events outside Europe**	**Broad dissemination of the project results outside Europe**	Track OpenScienceLink-related events to which Consortium members participate outside Europe	GO1, SO7	0	0	>=2	>=2
KPI27	**Number of marketing material** (leaflets, banners, fact sheets, posters, etc) **produced and distributed to promote the OpenScienceLink Services and pilots**	**Broad dissemination of the project results**	Track OpenScienceLink Service- and pilot-related marketing material produced and distributed by Consortium members	GO1, SO7	>=100	>=400	>=800	>=1000
KPI28	**Number of Targeted Contacts and Approached Potential Customers** (publishers, research organizations, universities)	**Development of a Rich Customer Base**	Track stakeholders and potential customers contacted by Consortium partners	GO1, SO7	>=5	>=15	>=25	>=40
KPI29	**Number of Policy Makers and Decision Makers** reached	**Development of a Rich Customer Base/Broad**	Track policy and decision makers	GO1, SO7	>=1	>=3	>=5	>=10

Ind. ID	KPI Name	Objective/ Expected Result	Method of Measurement	Related Project Objective(s)	Expected Progress			
					Year1	*Year2*	*Year3*	*Year4*[55]
		Exploitation Potential	contacted by Consortium partners					
Internal Business processes								
Core Business Achievements: *Improvement of Processes in Scientific Work's Publication, Sharing, Linking, Review and Evaluation*								
KPI30	**Speed of the Review Process**	**Improvement to the Review Process / Speed**	Use a control group performing the reviews in the conventional way / compare to OpenScienceLink	GO1, GO2, SO4	0	>=15%	>=30%	>=35%
KPI31	**Attraction of more competent reviewers**	**Improvement to the Review Process / Competence**	Use of Questionnaires	GO1, GO2, SO4	0	>=50%	>=100%	>=100%
KPI32	**Increased number of reviewers per review** with the same effort and in the same time frame	**Improvement to the Review Process / Number of Reviews**	Use a control group allocating reviews in the conventional way / compare to OpenScienceLink	GO1, GO2, SO4	0	>=50%	>=100%	>=100%
KPI33	**Trend Detection Accuracy**	**Improved Detection of Research Trends**	Use of Questionnaires	GO1, GO2, SO5	0	>=20%	>=50%	>=60%
KPI34	**Average increase of number of data papers creation per researcher**	Boost the **creation/production of data papers**	Keep track of statistics about data paper creation	GO1, GO2, SO3	0	>=15%	>=50%	>=70%
KPI35	**Correctness** of implicitly identified relationships between researchers and research groups (based on data mining, collaborative filtering etc.)	Boost **linking** and **collaboration** between researchers	**Percentage of recommendations which are relevant to the expert's topic/domain and are not part of his**	GO1, GO3, SO8	>=50%	>=65%	>=80%	>=80%

Ind. ID	KPI Name	Objective/ Expected Result	Method of Measurement	Related Project Objective(s)	Expected Progress			
					Year1	*Year2*	*Year3*	*Year4*[55]
			existing collaborations					
KPI36	Production of **metrics** that **improve** g-index, h-index, ISI for specific tasks - **Measurement of Improvement**	Produce **objective metrics** of **scientific performance**	Interviews with Experts / Questionnaires	GO1, GO2, SO5	0	>=10%	>=20%	>=30%
KPI37	**Number of distinct Best Practices (BPs) and Blueprints Produced**	**Produce/Promote Best Practices and Blueprints on Open Access**	Track the number of unique BPs	GO3, SO8	0	>=2	>=8	>=8
Learning & Growth perspective *Involvement of young researchers*								
KPI38	Number of young researchers in the project	Involvement of partners with great learning and growth potential	Track the young researchers (PhD students, postdoc) who participate in the project	SO7	>=10	>=10	>=10	-
KPI39	Number of young researchers involved in the pilots	Involvement of pilot users with great learning and growth potential	Track the young researchers (PhD students, postdoc) who participate in the pilots	SO7	0	>=100	>=400	>=1000

6 Summary and Conclusions

In this deliverable the requirements engineering processes associated with all stakeholders of the OpenScienceLink platform have been presented in detail. More specifically, within this document, the requirements associated with the interfacing to multiple openly accessible scientific repositories (including the repositories to be used during the pilot operations) were illustrated. In addition, the main use cases associated with the OpenScienceLink pilot services, along with detailed Key Performance Indicators (KPIs) for their monitoring and auditing were presented. In all, the current document summarized the OpenScienceLinik project landscape, and discussed the objectives of the project which set the basis for the OpenScienceLink platform requirements, and analysed what should be the requirements from all perspectives to be taken into account for the design and the implementation of the OpenScienceLink platform.

7 References

7.1 Research papers and studies

Alston, C., L. Paget, G. Halvorson, B. Novelli, J. Guest, P. McCabe, K. Hoffman, C. Koepke, M. Simon, S. Sutton, S. Okun, P. Wicks, T. Undem, V. Rohrbach, and I. Von Kohorn (2012). Communicating with patients on health care evidence. Discussion Paper, Institute of Medicine, Washington, DC. Retrieved from http://www.iom.edu/evidence.

Abdoul H., Perrey C., Amiel P., Tubach F., Gottot S., Durand-Zaleski I., Alberti C. (2012). Peer review of grant applications: criteria used and qualitative study of reviewer practices. PLoS One, 7(9), e46054.

Abdoul H., Perrey C., Tubach F., Amiel P., Durand-Zaleski I., Alberti C. (2012). Non-financial conflicts of interest in academic grant evaluation: a qualitative study of multiple stakeholders in France. PLoS One, 7(4), e35247

Bahill, A.T., & Gissing, B. (1998). Re-evaluating systems engineering concepts using systems thinking. *IEEE Transaction on Systems, Man and Cybernetics, Part C: Applications and Reviews, 28*(4), 516-527.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res., 28*(1), 45–48. http://dx.doi.org/10.1093/nar/28.1.45.

Ball, C.A., Dolinski, K., Dwight, S.S., Harris, M.A., Issel-Tarver, L., Kasarskis, A., Scafe, C.R., Sherlock, G., Binkley, G., Jin, H., et al. (2000). Integrating functional genomic information into the Saccharomyces Genome Database. *Nucleic Acids Res., 28*(1), 77–80. http://dx.doi.org/10.1093/nar/28.1.77.

Bar-Ilan, J. (2008) Which h-index? - A comparison of WoS, Scopus and Google Scholar, *Scientometrics*, 74(2), 257-271

Björk, B.C., & Hedlund, T. (2009). Two scenarios for how scholarly publishers could change their business model to open access. *J. Electronic Publishing, 12*, http://dx.doi.org/10.3998/3336451.0012.102.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST--database for “expressed sequence tags.” *Nature Genet., 4*(4), 332–333.

Bollen J, Van de Sompel H, Hagberg A, Chute R. (2009) A principal component analysis of 39 scientific impact measures. *PLoS One*, 4(6), e6022.

Bonetta, L. (2007). Scientists enter the blogosphere. *Cell, 129*(3), 443-445. Retrieved from www.sciencedirect.com/science/article/pii/S0092867407005430 [4 July, 2013].

Campbell, P. (2008) Escape from the impact factor. *Ethics in Science and Environmental Politics,* 8, 5–7.

Costanzo, M.C., Hogan, J.D., Cusick, M.E., Davis, B.P., Fancher, A.M., Hodges, P.E., Kondu, P., Lengieza, C., Lew-Smith, J.E., Lingner, C., et al. (2000). The Yeast Proteome Database (YPD) and Caenorhabditis elegans Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res., 28*(1), 73–76. http://dx.doi.org/10.1093/nar/28.1.73.

Dennis, A., Wixom, B.H., & Tegarden, D. (2005). *Systems Analysis and Design with UML Version 2.0: An Object-Oriented Approach*, 2nd Edition. Hoboken, NJ: John Wiley & Sons.

Dewatripont M., et. al. (2006). European Commission, Directorate-General for Research, Directorate C Science and Society, Unit C2 Scientific Advice and Governance. Study of the Economical and technical evolution of publication markets in Europe. (final report 2006).

Retrieved from http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.

Gideon Emcee, C. (2009). Building a sustainable framework for open access through information and communication technologies, IDRC-CRDI, 41. Retrieved from http://idl-bnc.idrc.ca/dspace/handle/10625/41336.

Gilbert, F., & Ovadia, D. (2011). Deep brain stimulation in the media: over-optimistic portrayals call for a new strategy involving journalists and scientists in ethical debate. *Frontiers in Integrative Neuroscience, 5*(16). Retrieved from www.ncbi.nlm.nih.gov/pmc/articles/PMC3095813/ [4 July, 2013].

Guibault, L. (2013). Licensing research data under Open Access conditions, in D. Beldiman (ed.), *Information and knowledge: 21st century Challenges in Intellectual property and knowledge governance*. Cheltenham, Edward Elgar.

Gurabardhi, A., Gutteling, J.M., & Kuttschreuter, M. (2004). The development of risk communication: an empirical analysis of the literature in the field. *Science Communication, 25*(4), 323-349.

Harnad, S., & Brody, T. (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine 10*(6). Retrieved from www.dlib.org/dlib/june04/harnad/06harnad.html.

Harnad, S., et al. (2008). The access/impact problem and the green and gold roads to open access: an update. *Serials Review 34*(3).

Hooks, I. (2005). *Writing Defect-Free Requirements*. Boerne, TX: Compliance Automation.

Houghton, J.W., & Sheehan, P.J. (2009). Estimating the potential impacts of open access to research findings. *Economic Analysis and Policy, 39*. Retrieved from http://www.eap-journal.com.

Hugenholtz P.B., Van Eechoud M., Van Gompel S. et al. (2006). Recasting of Copyright and Related Rights for the Knowledge Economy, study prepared for the European Commission ETD/2005/IM/D1/95, Amsterdam. Available at: http://www.ivir.nl.

INCOSE (2006). *The SIMILAR Process, A Consensus of the INCOSE Fellows*. Retrieved from www.incose.org/practice/fellowsconsensus.aspx.

Janssen, K., & Dumortier, J. (2006). The protection of maps and spatial databases in Europe an the US by copyright and the sui generis right, *The John Marshall Journal of Computer and Information Law, 2*, 195-226.

Kirby, T. (2011). Science media centres go global. *The Lancet, 377*, 285. Retrieved from www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)60078-0/fulltext [4 July, 2013].

Laakso M, Welling P, Bukvova H, Nyman L, Björk B-C, et al. (2011) The Development of Open Access Journal Publishing from 1993 to 2009. *PLoS ONE,* 6(6), e20961.

Ladle, R. J., Jepson, P., & Whittaker, R. T. (2005). Scientists and the media: the struggle for legitimacy in climate change and conservation science. *Interdisciplinary Science Review, 30*(3), 231-240.

Lery, T., Bressler, P. (2007). Earnest Report on Researchers' Requirements, TERENA. Funded by the European Community through the GN2 project in the Sixth Framework Programme for Research and Technological Development.

Lewis, D.W. (2012). The Inevitability of Open Access. *College & Research Libraries, 73*(5), 493-506. Retrieved from http://crl.acrl.org/content/73/5/493.full.pdf.

Lohr, K.N. (2004). Rating the strength of scientific evidence: relevance for quality improvement programs. *International Journal for Quality in Health Care, 16*(1), 9–18.

Max Plank Society (2013). Signatories. Open Access at the Max Plank Society. Retrieved from http://oa.mpg.de/lang/en-uk/berlin-prozess/signatoren.

MITRE (2012a). Eliciting, Collecting, and Developing Requirements, in *Systems Engineering Guide.* McLean, VA: MITRE. Retieved from www.mitre.org/work/systems_engineering/guide/se_lifecycle_building_blocks/requirements_engineering/eliciting_collecting_developing_requirements.html.

MITRE (2012b). Analyzing and Defining Requirements, in *Systems Engineering Guide.* McLean, VA: MITRE. Retieved from www.mitre.org/work/systems_engineering/guide/se_lifecycle_building_blocks/requirements_engineering/analyzing_defining_requirements.html.

Neill, U.S. (2008). Publish or perish, but at what cost? *J Clin Invest., 118*. 236.

Nightingale J.M., Marshall G. (2012) Citation analysis as a measure of article quality, journal influence and individual researcher performance. *Radiography* 18 60e67.

Pendlebury, D.A. (2008). Using Bibliometrics in Evaluating Research. Research Department, Thomson Reuters, Philadelphia, PA.

Piwowar, H.A., Daym R.S., and Fridsma, D.B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE 2*(3): e308. http://dx.doi.org/10.1371/journal.pone.0000308.

Reed, R. (2001). (Un-)Professional discourse journalists' and scientists' stories about science in the media. *Journalism, 2*(3), 279-298.

Reich E.S., Myhrvold C.L. (2013). Funding agencies urged to check for duplicate grants. Nature 493(7434), 588-9.

Reinhold, W.C., Sunshine, M., Liu, H., Varma, S., Kohn, K.W., Morris, J., Doroshow, J., and Pommier, Y. (2012) CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. *Cancer Res* 72(14); 3499–511. http://dx.doi.org/10.1158/0008-5472.CAN-12-1370.

San Francisco Declaration on Research Assessment (2012). Annual Meeting of The American Society for Cell Biology (ASCB) in San Francisco, CA.

Schroter S., Groves T., Højgaard L. (2010). Surveys of current status in biomedical science grant review: funding organisations' and grant reviewers' perspectives. BMC Med. 8:62.

Schwitzer, G. (2008). How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. *PLoS Med, 5*(95). Retrieved from www.ncbi.nlm.nih.gov/pmc/articles/PMC2689661/ [4 July, 2013].

Solomon, D.J. (2008). *Developing Open Access Journals: A practical guide*. Oxford, UK: Chandos Publishing.

Surfdirect (2009). The legal status of raw data: a guide for research practice. CIER. 52.

Van Eechoud, M. (2012). Along the road to uniformity- diverse readings of the Court of Justice Judgements on copyright work, JIPITEC 1 –2012, 60-80.

Vaughan, L.; Shaw, D. (2008) A new look at evidence of scholarly citations in citation indexes and from web sources, *Scientometrics*, 74(2), 317-330.

Wagner, J. A., E. C. Wright, M. M. Ennis, M. Prince, J. Kochan, D. J. Nunez, B. Schneider, M. D. Wang, Y. Chen, S. Ghosh, B. J. Musser, and M. T. Vassileva (2009). Utility of adiponectin as a biomarker predictive of glycemic efficacy is demonstrated by collaborative pooling of data from clinical trials conducted by multiple sponsors. *Clinical Pharmacology & Therapeutics* 86(6): 619-625.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., and Rapp, B.A. (2000). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res., 28*(1), 10–14. http://dx.doi.org/10.1093/nar/28.1.10.

Zarin, D. (2012). Statement of the Director of ClinicalTrials.gov at the U.S. National Library of Medicine, National Institutes of Health (NIH) during the Workshop on Sharing Clinical Research Data, October 4–5, 2012, U.S. National Academy of Sciences, Washington D.C.

Zarin, D. A., T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide (2011). The ClinicalTrials.gov results database – update and key issues. *New England Journal of Medicine 364*(9): 852-860.

7.2 Declarations and policies

Berlin Declaration (2003). Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, 22 October 2003. Retrieved from http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf.

Directory of Open Access Journals (2013, June). *DOAJ announces new selection criteria*. Retrieved from www.doaj.org/doaj?func=news&nId=303.

DOW (2012). *Open Semantically-enabled, Social-aware Access to Scientific Data*, Grant agreement no. 325101. Annex I - Description of Work. OpenScienceLink consortium.

EUROHORC (2008). Recommendations on Open Access. Retrieved from www.eurohorcs.org/SiteCollectionDocuments/EUROHORCs_Recommendations_OpenAccess_200805.pdf.

Oregon State University (2013). OSU adopts university-wide open access policy. *OSU News & Research Communications*. Retrieved from http://oregonstate.edu/ua/ncs/archives/2013/jun/osu-adopts-university-wide-open-access-policy.

PILA (2012). *Membership Qualifications and Rules*, Publishers International Linking Association, as updated on 17 September 2012. Retrieved from www.crossref.org/02publishers/59pub_rules.html.

San Francisco Declaration (2012). San Francisco Declaration on Research Assessment. Annual Meeting of The American Society for Cell Biology (ASCB), San Francisco, CA.

7.3 EU documents

Article 29 Working Party, Opinion on the concepts of 'controller' and 'processer', 35.

Article 29 Working Party, Opinion on concept of personal data, 26.

Charter of the Fundamental Rights of the European Union (2000/C 364/01), www.europarl.europa.eu/charter/pdf/text_en.pdf.

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions for a reinforced European Research Area Partnership for Excellence and Growth, Brussels, COM (2012) 392 final.

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions towards better access to scientific information: boosting the benefits of public investments in research, Brussels, COM (2012) 401 final.

Commission recommendation of 17.7.2012 on access to and preservation of scientific information, Brussels, COM (2012) 4890 final.

Directive 95/46/EC of the European Parliament and of the Council of 24.10.1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive), OJ L 281.

European Commission, Proposal for a Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, instigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data, Brussels, 25 January 2012, COM (2012) 10 final.

European Commission, Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Brussels, 25 January 2012, COM (2012), 11 final.

European Commission (2013). Digital science in Horizon 2020, Concept Paper. Brussels: European Commission, Digital Agenda for Europe. Retrieved from:

http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=2124.

European Convention of Human Rights. Retrieved from:

www.echr.coe.int/Documents/Convention_ENG.pdf.

7.4 Case Law: European Court of Justice

C-203/02, British Horseracing board Ltd, 2004

C-388/02, Fixtures Marketing, 2004.

C-304/07, *Directmedia/Albert-Ludwigs-Universität Freiburg,* 9 October 2008.

C-5/08, *Infopaq International A/S v Danske Dagblades Forening,* 16 July 2009.

C-393/09, *Bezpecnostni softwarova asociace v. Ministerstvo kultury,* 22 December 2010.

C-145/10, *Eva maria Painer v. Standard Verlag GmbH,* 1 December 2011.

C-604/10, *Football Dataco v. Yahoo UK Ltd.,* 1 March 2012.

7.5 OpenScienceLink Documents

OpenScienceLink Consortium. (2013). *OpenScienceLink: OpenSemantically-enabled, Social-aware Access to Scientific Data.* EC.