Technical background notes for Framework Programme 7, Strategic Objective ICT-2013.4.3 "SME initiative on analytics"

DG CONNECT/G3
CNECT-G3@ec.europa.eu
http://cordis.europa.eu/fp7/ict/content-knowledge/home_en.html

This document is intended to provide background information and technical commentary on Strategic Objective ICT-2013.4.3 of Framework Programme 7 (FP7). The official text of the objective (including the timeline and procedures for applications) has been published as part of the 2013 FP7 work programme

http://cordis.europa.eu/fp7/ict/docs/ict-wp2013-10-7-2013.pdf

The official text is the only legally binding source of information on the strategic objective. Should any inconsistency between the present explanatory document and the official text be detected it is always to be resolved in favour of the work programme text.

Motivation of the objective and scope of this document

The official text of the work programme states that this objective concentrates on:

"Helping European Small and Medium Enterprises acquire the competences and resources they need to develop innovative content and data analytics services." (page 51)

The motivation behind this objective is to support European Small and Medium Enterprise in their innovation efforts involving language or data technologies. In both cases the emphasis is not so much in advanced, long term, research as much as in the judicious deployment in practically motivated situations of novel techniques and/or resources that up until now have not been tested in the field.

a) Integrated Open Data Incubator

Development of services based on the use of available data, particularly from public bodies, is specifically required for theme a) of the objective, which reads:

An Integrated Project to establish an environment and calling for efficient, small scale development of services of commercial interest based on the use of European open data by Small and Medium Enterprises (SMEs). The IP should:

- *devote most of its resources to publish and manage regularly scheduled and well-advertised calls for SMEs to submit mini-proposals to be funded for a period between six and twelve months.*

- *create a computing infrastructure where the winning mini-proposals will find accurate, up-to-date and (when useful and feasible) linked versions of the data they need for their services and, if they so wish, deploy the experimental version of their services.*

- *establish a mechanism for connecting open data demand and supply by systematically contacting European public bodies for their open data and assisting them in the efficient and sustainable publication of such data, if needed with targeted engagements.*

- *solicit open data reuse ideas from the general public and conduct a European wide open data reuse information campaign.*

- *The IP will finally create a process to connect the most successful SMEs with sources of funding and business networks.*

Here we provide additional details on the mechanics of the exercise and some guidance on some design principles that should inform proposals (and later, projects) responding to this part of the call.

What is the Open Data Reuse Incubator (ODRI)?

The objective of this call is to make sure that technical obstacles to the reuse of European open data are removed to the extent possible.

While a growing number of regional, national and international administrations (the European Commission among them) are making available growing amounts of open data, the number of applications (and the number of users of those applications) based on the reuse of those data have arguably not been growing at the same rate, despite excellent national initiatives in this direction.[1]

It is thus possible that open data of good quality and potential interest to some end user goes unutilised due to some obstacle of administrative or technical nature. ODRI is designed to create an environment in which those obstacles are systematically removed and Small and Medium Enterprises (SME) are offered various forms of support (including financial) to pursue their data reuse ideas.

In this respect, ODRI's high level objectives are quite similar to those of the recently launched UK's Open Data Institute[2] (ODI) with the difference that while ODI is focused on promoting the reuse of UK open data, ODRI does the same for data from all over the EU.

How will ODRI help logistically?

The objective of ODRI is to allow SMEs to concentrate on developing and testing their data based application. Thus ODRI will be tasked with building and maintaining data and infrastructure for SMEs to deploy their prototypes.

An SME approved for work within ODRI will find at ODRI all the data they need to develop their application plus a computing environment where they could deploy their prototype application for testing.

[1] http://www.etalab.gouv.fr/m/article-105550957.html http://www.appsforitaly.org/en/

[2] http://theodi.org

If the data required by the SME is known to exist but is not yet available at ODRI, it will be ODRI's task to obtain it from the publisher and perform whatever reasonable transformations (e.g. format conversions, linking to other datasets) are required for the SME to proceed with their idea.

In order to 'prime the pump' of the data available on its platform, ODRI will engage the relevant actors on both the data demand and supply directions:

- on the demand side, ODRI will carry out a EU-wide information campaign extracting from the public what kind of applications they would like to see developed based on EU open data;

- on the supply side, ODRI will engage with public administrations all over the EU inquiring about their plans and schedule for the release of datasets, sharing with said administrations information collected at the point above on what type of data the public perceives as most useful in order to develop applications. If an administration is in principle willing to publish desired datasets but considers that it doesn't have the technical know-how to do so proficiently, ODRI will offer its technical assistance in the context of a limited intervention. ODRI personnel would visit the administration and assist with the publication task. The ODRI will also assess the technical maturity of the administration and provide guidance (in the form of deployment of and hands-on training for open source data management software whenever this is possible) on how the administration could proceed to publishing future datasets by itself

How will ODRI help SMEs get in and out of the ODRI environment?

ODRI will be required to assist SMEs in being admitted to the ODRI environment and in planning their exit.

On the admission end, it will be ODRI's responsibility to advertise calls for proposal submissions by SMEs. This means that it will be ODRI's responsibility to use social networks, conferences and any other relevant means in order to spread its message where promising software development SMEs (even some who might not naturally have thought of themselves as open data re-users) are naturally found

On the exit end, it will be ODRI's responsibility to educate its graduate SMEs on product development strategies and connect them with previous graduates but, most importantly, additional opportunities for funding, from venture capital to regional development funds. It is thus important that ODRI build and grow a solid network of relevant contacts.

How will ODRI function in practice?

ODRI will formally exist as an Integrated Project, a project funded as a result of a competitive call for proposals submitted by a consortium satisfying the requirements for participation of Framework Programme 7.

Such proposals will not be about building applications based on the reuse of open data but rather about building the ODRI environment in which others (and specifically SMEs) will do so. Proposals will be evaluated based on the credibility of the consortium's plans for providing all the forms of support described above.

The consortium will spend only a fraction of the funds it receives (the budget for the call is 5 million Euros) to set up and operate the ODRI environment.

The consortium will disburse the majority of the funds to fund mini-proposals of a duration between six and twelve months submitted by SMEs in response to broadly advertised calls. Based on the amount of funding commonly offered by seed venture capital for efforts of this scope, it is expected that funding for each such mini-proposal could range between 50,000 and 150,000 Euros

b) Easing transfer and take-up of language technologies

Language technologies are often deployed within products and services relating to web or enterprise intelligence, including text and audio mining, social media analytics and sentiment analysis, enterprise search and content management, online and cloud based translation, etc.

This action targets focused user- and market-oriented projects in any of the above areas, with the overall goal of bringing language technologies closer to commercial maturity through an "industrialisation" process including but not limited to: i) engineering of promising but commercially untried technologies, e.g. in terms of performance, robustness and coverage; ii) integration within existing or upcoming products and services; iii) first-use experimentation and validation in a clearly identified application domain; iv) in-depth assessment along technical, used related and economic dimensions; (v) identification of possible exploitation paths and viable business models, and of suitable sources of funding.

The idea behind outcome b) is to demonstrate practical usefulness of language technologies in real-life applications and to stimulate and encourage take-up of language technologies, especially by SMEs that may not have been applying them so far. In other words, technology transfer is an important element of this outcome. In order to credibly demonstrate integration of language technologies in various business processes, products and services, the project should set as its target to produce, within the duration of the project, a live and functional outcome (system, service, concept) with real users. While this "outcome" does not have to be a full-fledged product and need not have all the intended final functionalities and features, it should at least provide a "beta" which is mature enough to usefully demonstrate the results in a (limited) real-life setting.

As the duration of the projects will be relatively short, this means that the language (and other) technologies to be applied need to be relatively mature. There is very little (if any) room for research, some room for development, but experimenting, evaluating and validation are essential.

Projects should arise from the real business cases of their (SME) partners rather than artificially constructed scenarios. "Users" should be the actual clients and stakeholders of the project partners. However, a good proposal will identify emerging trends in markets and society, and will address them proactively.

Impact is the most important success criterion for the proposed action. Therefore, the proposal should clearly demonstrate (preferably with facts and figures) what the impact would be. Look at your project idea from an outsider's perspective and ask yourself what difference you will make, which problems you will solve, how you will improve existing systems and processes and who would benefit from it. Impact can come in many different ways. It can be direct expansion of business and winning new markets. But it can also come indirectly through provision of highly useful open-source tools to the SME community, enabling them to win new markets, lower their costs or streamline their processes.

Finally, a successful proposal proposes something that is new. A mere extension of an existing system or service, or a duplication or "enhanced version" of a competing service is not likely to rank high in this call. However, it is necessary to strike a reasonable balance between novelty and practicality. Novelty does not mean science fiction or unrealistic, artificial scenarios, it can mean new ways of winning customers, making business, creating new markets by innovative, creative and cost-effective ways of integrating language technologies and automation into workflows and business models.

c) Software components and intuitive end user applications based on reuse of open data

Outcome c) of the objective calls for

Development of software components supporting the whole life cycle of reuse of multilingual open data, particularly from public bodies.

The vision behind this objective is to make the reuse of open data easy and effective on the part of SMEs. This requires creating a process and software stack where all the most common steps in the publishing and reusing of data are addressed from the perspective of the practitioner.

In this context, there are several groups of practitioners or domains whose needs have to be understood and addressed.

A first group of practitioners are the people who (typically but not necessarily in public bodies), are charged with the task of publishing open data. These people need support to connect their current processes with processes that will create sustainable open data publishing operations. A good analogy of what is expected in support of this group of users is the set of tools that emerged in the mid-90s to support the publication of well-formed HTML documents by non-technical users who were primarily interested in the content of documents to be published on the web. We are today in a similar situation, with many data publishers who have deep expertise in the subject matter of the data they wish to publish but little or no expertise in the techniques to do so effectively in a web environment. Another important need to be met is the development of usable tools for data curation (removing duplicates, using uniform scales for dimensional information, identifying outliers, etc…). Examples of such tools in existence today are Refine[3] (and its Linked Data/RDF extensions), Data Wrangler[4] but there might be data practices niches that deserve specialised tools. It is in support of this

[3] http://code.google.com/p/google-refine/

[4] http://vis.stanford.edu/wrangler/

population of users that we invite the development of data publishing tools that are very easy to learn and to use and very simple to deploy and maintain in organisational IT environments that may be restrictive and technically not very sophisticated. These are all properties that are indispensable to guarantee the sustainability of efficient data publishing operations. In addition, the tools to be developed will have to support data publications best practices in ways that do not conflict with the existing processes of an organisation and that allow the organisation itself to take maximum advantage of the data it publishes.

A second sub-objective is the development of easy to deploy and easy to use tools for the interlinking of datasets, particularly in a web environment. An important aspect that these tools will need to address convincingly is the issue of the reuse of ontologies and entity identifiers. Once again, all these functionalities will need to be delivered assuming the typical user will be a data publisher or a data re-user, i.e. individuals who use ontologies and identifiers as a means for accomplishing some externally motivated goal and not as an object of scientific studies. The tools to be developed will need to include the means to determine how well those goals are being met and to what extent data curation and linking improves the performance towards those goals over time.

A third sub-objective is the development of tools designed to capture user feedback and use the feedback to optimise the behaviour of applications based on open data. As a hypothetical example, an open data reuse application might notice that users often ask for transportation information in a region where this kind of open data is not available. Instead of requesting this information from the relevant public body on general principles, the application developer could then show detailed statistics to prove how often these data would be useful in her application. Similarly, users might routinely go through a number of steps in the application that betray an information intent that is not currently supported. In this case the developer should be using these statistics as a guide for the type of functionalities she should be developing in support of her users. The collection and interpretation of these statistics is not always trivial and the existence of such tools could greatly simplify the efforts of a data re-use developer.

Finally, also invited are tools to ease the cross-platform development of data applications: while many existing data re-use applications are clearly conceived for the needs of the desktop analyst, the very rapid spread of smartphones and tablets makes it likely that there will be many mobile applications whose functionalities users may wish to have augmented by means of open data. Indeed, this may apply also to applications not aimed at humans as would be the case for automation or robotic platforms. The ability of data re-use developers to develop data-based applications and deploy them effortlessly across many different computing platforms will greatly benefit from the development of data aware application development toolkits appropriate for the working environments of SMEs.

Appendix: a list of questions that proposals must answer in order to fulfil strategic objective ICT-2013.4.3

This appendix contains a list of simple questions that a consortium should ask about the proposal to be submitted. If the proposal as submitted does not contain a clear answer to the majority of the relevant questions for the various outcomes it places itself at a serious disadvantage in a very competitive selection process.

a) Integrated Open Data Incubator

1. How is the consortium putting the needs of SMEs (as opposed to those of the core partners) at the centre of its planning?
2. How is the consortium planning to maximise the amount of resources available to SMEs and minimise the amount used by the core partners of the consortium?
3. How is the consortium planning to attract promising SMEs, assist them in the process of submitting a proposal to the planned calls and assist the select SMEs in complying with the requirements of participation in a FP7 project?
4. Does the consortium have a strategy concerning the data resources that it plans to curate for the benefit of SMEs? Are there data domains that mutually increase each other's value if well curated?
5. Does the consortium have clear plans (and credible cost projections) for the computational infrastructure that it plans to make available to SMEs for the development of application prototypes?
6. Does the consortium have clear plans to help SMEs migrate from the prototype environment to independent applications?
7. Does the consortium have a clustering strategy to encourage the participation of SMEs with mutually value enhancing application ideas?
8. What is the consortium's specific strategy to connect SMEs with additional sources of support (e.g. venture capital)? Does the consortium have quantitative targets in this respect? If so, how will it monitor its progress?

b) Easing transfer and take-up of Language Technologies

1. What is the (real-life) problem that you want to solve?
2. What difference will your solution make and how broad and large will the impact be?
3. What tangible results will your project deliver?
4. What is the level of maturity of the language technologies to be applied in the project?
5. Which languages will you address?
6. How will the service be sustained after the end of the project? What is the sustainability/revenue model, sources of income or financing?
7. Who are the users and stakeholders of your project? How do you intend to reach, mobilize and convince them?

c) Software components and intuitive user applications based on reuse of open data

1. Why is the component proposed needed? Who, specifically, has a need for it and why this need cannot be satisfied by data processing components already available?

2. What is the specific functionality that will be offered by the component? What would demonstrate that the component's functionalities have or have not been implemented successfully?
3. How will the project address the issue of usability (both from the point of view of ease of installation and maintenance and from the point of view of end user ease of learning and operation)? What specific process will be put in place to constantly monitor and improve usability?
4. On which platform(s) is the proposed component intended to be deployed? How will the consortium prove that the deployment is successful?
5. If the proposed component is not a complete application, how will the consortium prove that it will be beneficial for build (or as a part of) complete data applications? How will the consortium prove that the proposed component will be easy to integrate with other existing or future data processing components?