OKKAM

Enabling the Web of Entities

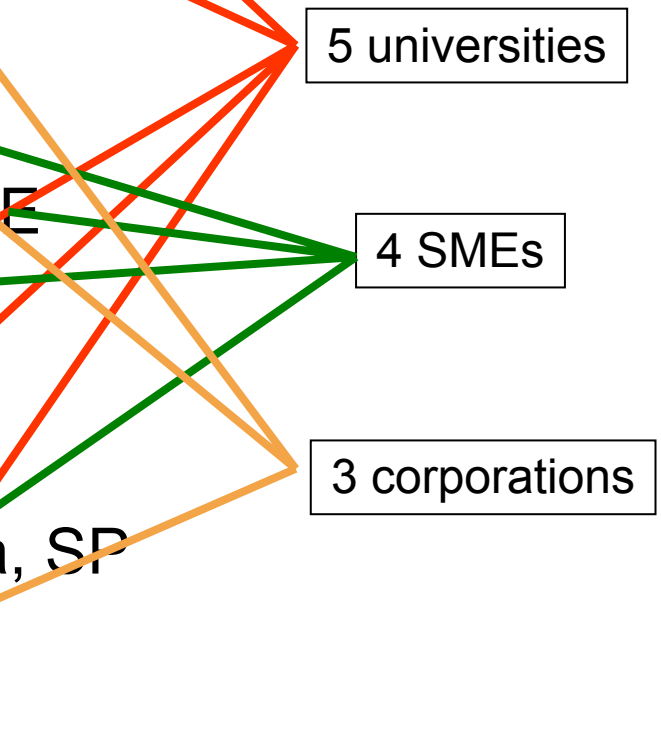
Large-Scale Integrating Project (GA#215032)

The OKKAM factsheet

Project Title	**Enabling the Web of Entities**. A scalable and sustainable solution for systematic and global identifier reuse in decentralized information environments
Acronym	**OKKAM**
Starting date	01/01/2008
End date	30/06/2010
Duration	30 months
No. of partners	12 (5 universities, 4 SMEs, 3 large companies)
Effort	763 person months (more that 63 person years)
Total cost	7.352.931,34 Euro
EU contribution	5.125.000,00 Euro
Type of instrument	Large-scale Integrating Project

The OKKAM consortium

1. University of Trento, IT
2. L3S Hannover, DE
3. SAP Research, DE
4. Elsevier, NL
5. Expert System, IT
6. Europe Unlimited, BE
7. MAC, IE
8. EPFL, CH
9. DERI Galway, IE
10. University of Malaga, SP
11. INMARK, SP
12. ANSA, IT

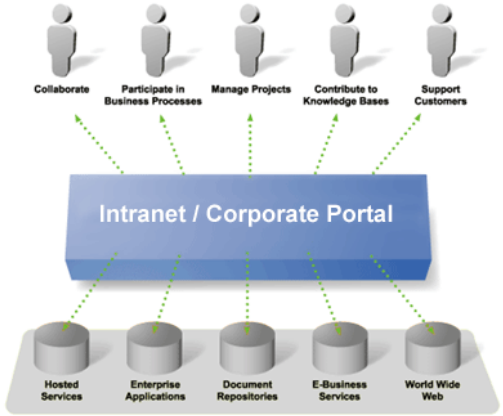


OKKAM: the rational

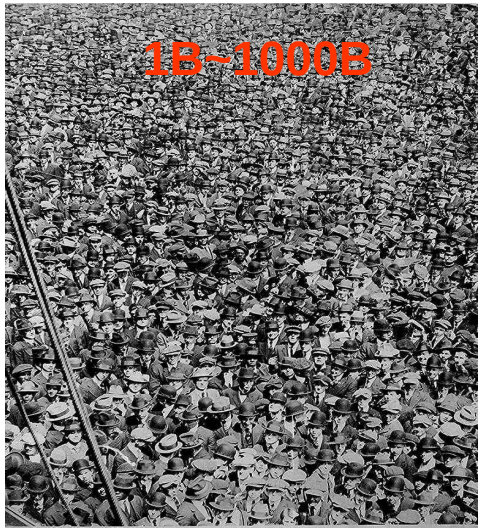
How many "entities" (persons, locations, organizations, events, projects, products, …) are named in:

Files in your laptop

Content in your Intranet/ Enterprise Information System

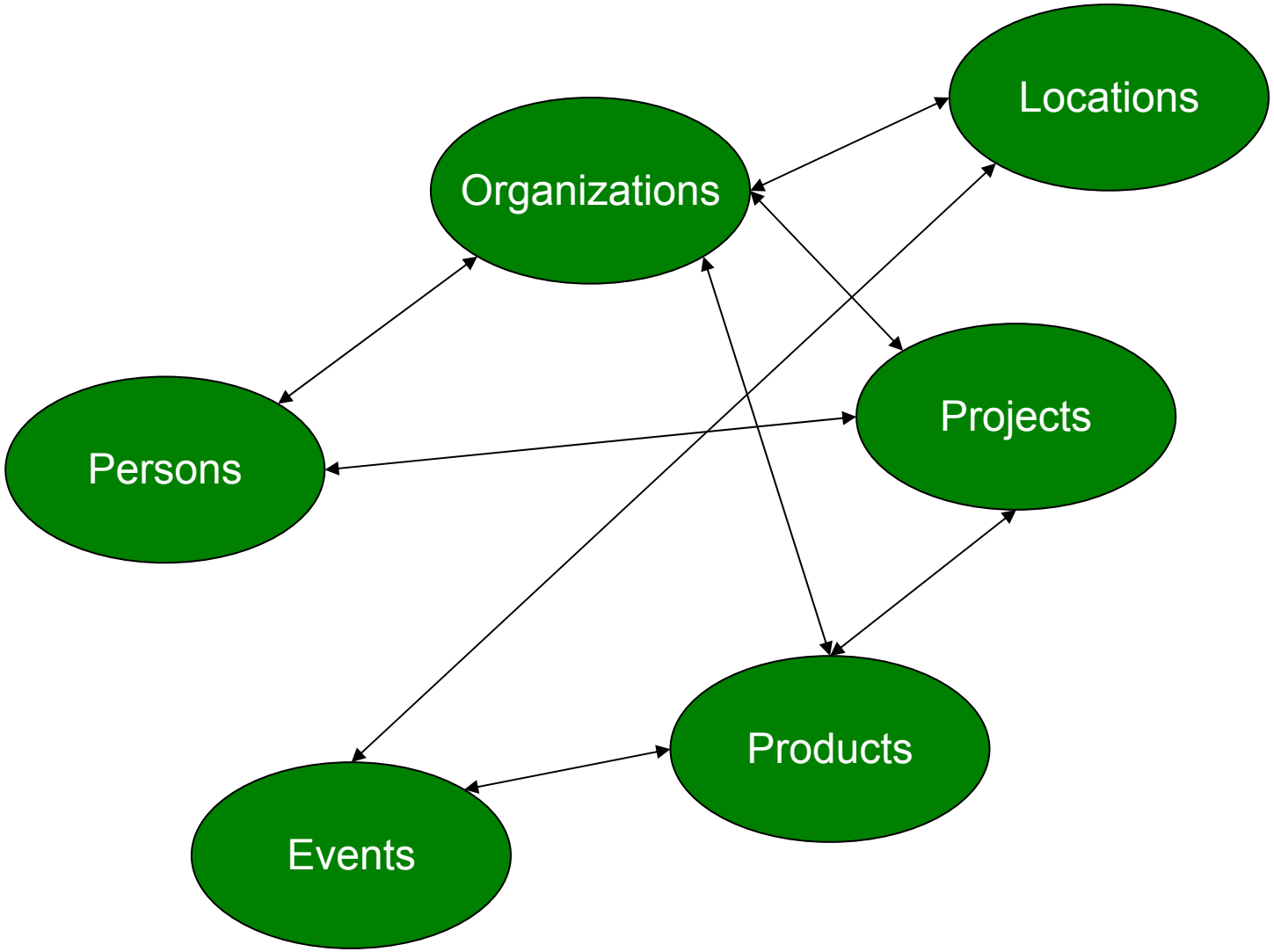


Contents/applications on the Web



An invaluable asset

"Entities" is what a large part of our knowledge is about:



However …

How many names, descriptions or IDs (URIs) are used for the same "entity"?

London 런던 ᱞᱚᱱᱰᱚᱱ ಲಂ ಡನ ಲಂ दन લં ડન □□□□ ロンドン লন্ডন ลอนดอน இலண்ட டன் ლონდონი Llundain Londain Londe Londen Londen Londen Londinium London Londona Londonas Londoni Londono Londra Londres Londrez Londyn Lontoo Loundres Luân Đôn Lunden Lundúnir Lunnainn Lunnon לאנדאן לונדון لندن لندن لوندون Λονδίνο Лёндан Лондан Лондон Лондон Лондон Լոնդոն 伦敦 …

The capital of UK, host city of the IV Olympic Games, host city of the XIV Olympic Games, future host of the XXX Olympic Games, the city of the Westminster Abbey, the city of the London Eye, the city described by Charles Dickens in his novels, …

http://sws.geonames.org/2643743/
http://en.wikipedia.org/wiki/London
http://dbpedia.org/resource/Category:London
…

… or …

How many "entities" have the same name?

London, KY
London, Laurel, KY
London, OH
London, Madison, OH
London, AR
London, Pope, AR
London, TX
London, Kimble, TX
London, MO
London, MO
London, London, MI
London, London, Monroe, MI
London, Uninc Conecuh County, AL
London, Uninc Conecuh County, Conecuh, AL
London, Uninc Shelby County, IN
London, Uninc Shelby County, Shelby, IN
London, Deerfield, WI
London, Deerfield, Dane, WI
London, Uninc Freeborn County, MN
...

- Or
  - London, Jack
    2612 Almes Dr
    Montgomery, AL
    (334) 272-7005
  - London, Jack R
    2511 Winchester Rd
    Montgomery, AL 36106-3327
    (334) 272-7005
  - London, Jack
    1222 Whitetail Trl
    Van Buren, AR 72956-7368
    (479) 474-4136
  - London, Jack
    7400 Vista Del Mar Ave
    La Jolla, CA 92037-4954
    (858) 456-1850
  - ...

… or …

How many content types / applications provide valuable information about each of these "entities"?

News about London

REUTERS

Revyu.com reviews on hotels in London

Wiki pages about the London

Pictures and tags about London

flickr

LinkedIn

Social networks in London

YouTube
Broadcast Yourself

Videos and tags for London

This is an immense loss of value

- Precision and recall of keyword-based **search engines** is deeply affected
- **Semantic search** still mainly relies on keyword search/matching (as very few URIs are consistently reused across RDF datasets)
- **Information integration** (e.g. from heterogeneous databases) is hard to achieve
- **Mash-ups** from different sources are not easy to create
- **Data/Text/Web mining** tools must struggle with object consolidation from different data sources
- **Information extraction** produces poorly integrated results (co-reference is supported mainly in the same document/corpus, not across large-scale distributed environments)

The OKKAM vision

OKKAM envisions an open, decentralized and global knowledge space in which:

- Every entity (individual, instance, “thing”) is assigned a global identifier, ideally unique
- The same identifier is used to name the same entity across any type of content, format, application, domain, language, culture
- Contents on the Web (from text to multimedia) are consistently annotated with entity unique identifiers
- Users and application are provided with easy and straightforward ways (e.g. GUIs, APIs) for retrieving the IDs of the entities they need to name in their content

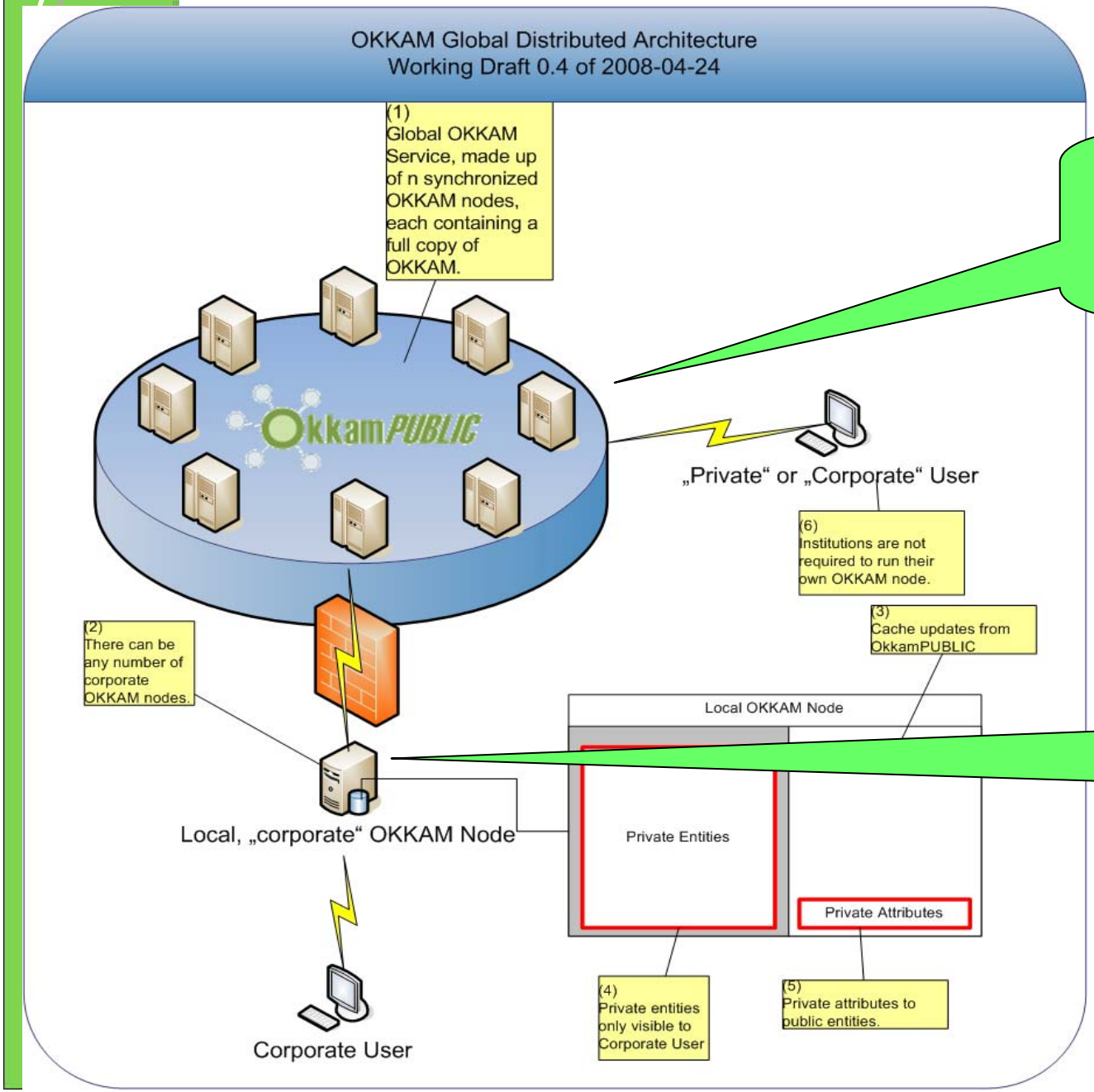
This will enable a new powerful generation of **entity-centric applications and services**

Enabling the OKKAM vision: an ENS for the Web

OKKAM aims at implementing an **Entity Name System** (ENS) for the Web, namely an *infrastructure* which can offer the following basic *services*:

- **ID storage and management**: stores, maintains and makes available for reuse IDs (URIs) for anything which is named in a networked environment
- **Entity matching (and look-up)**: maps any arbitrary description of an entity to its global ID (or to the collection of known IDs)
- **APIs**: provides application interfaces which can return IDs to applications which needs them in the creation of new content (e.g. word processors, ontology editors, HTML editors, web-based authoring environments, …)
- **Access**: allow users or application to create new IDs for things which do not have one, or to modify/update an entity description in the ENS

OKKAM ENS – Global and Decentralized



Replicated public nodes (ENS servers) hosted in different locations

Local “corporate“ nodes for non-public data (and cache)

ENS-empowered tools

Any application which is used to create/edit content can be extended with “plugins” which enable the application to get from the ENS the right ID for a named entity

Examples of OKKAM-empowered tools:

Annotating texts created with MS Word with IDs automatically obtained from the ENS

Reusing URIs in RDF/OWL knowledge bases built with Protégé with IDs automatically obtained from the ENS

Introducing global IDs to content created through web-based authoring interfaces (e.g. FOAF content)

Three OKKAM applications

In the project, we have three application scenarios:

- Entity-centric **semantic search engine** (DERI Galway)
- Entitity-centric **organizational knowledge management systems**: entity-centric product lifecycle management within SAP
- Multimedia **authoring environments**: supporting the creation of news (ANSA) and scientific papers (Elsevier) in an ENS-empowered environment

Organization of Work - Dimensions

- **Work Packages (Work breakdown structure)**
- **Project Competence Areas (PCAs)**
- **Integration and Demonstration Subprojects (ISP)**
- **Entity-centric Applications**

Work Breakdown Structure

15 Work Packages

- Further broken down into Tasks
- Result in Deliverables
  - Integrated SW Components
  - Documents (published, unpublished)

- Are not mapped 1:1 to a single partner
  - but are in the responsibility of a single partner
  - require task- or subtask-level co-operation

Project Competence Areas (PCAs)

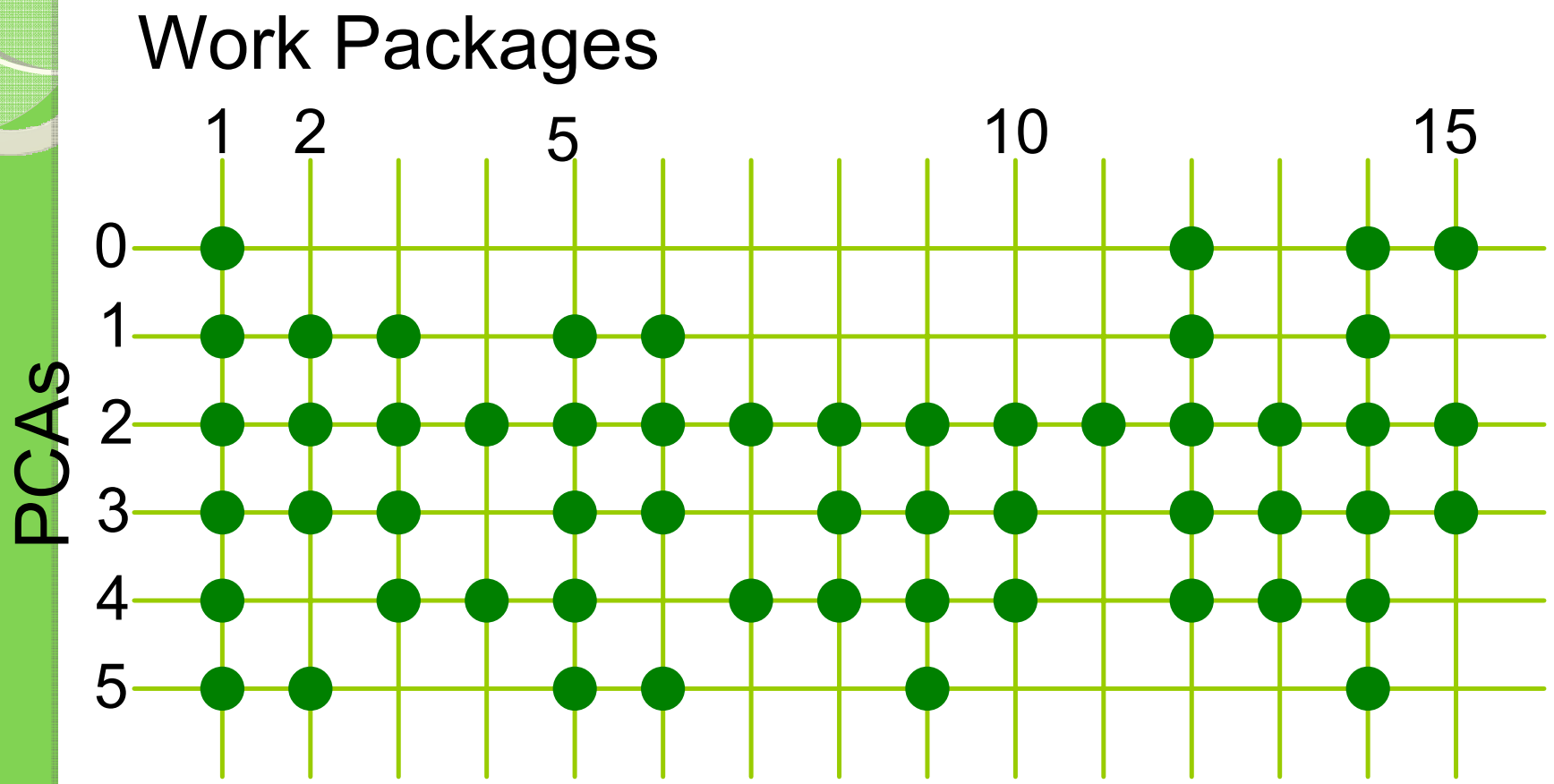
- Bundle related RTD competences
- Co-ordinates between related activities
- Fosters coherent development of contributions
- Ensures transition of research results into implementation

- Are a crucial instrument for the tight integration among partners and activities

Project Competence Areas

	Project Competence Area (PCA)	Area Leader
PCA-0	Community building, sustainability and exploitation	Europe Unlimited
PCA-1	Distributed storage & scalable entity management	EPFL
PCA-2	Entity representation, matching and reasoning	Univ. Trento
PCA-3	Entity lifecycle management and evolution	L3S Hannover
PCA-4	Information extraction, crawling and harvesting	ExpertSystem
PCA-5	Security, trust and privacy	Univ. Malaga

PCA vs. Work Packages

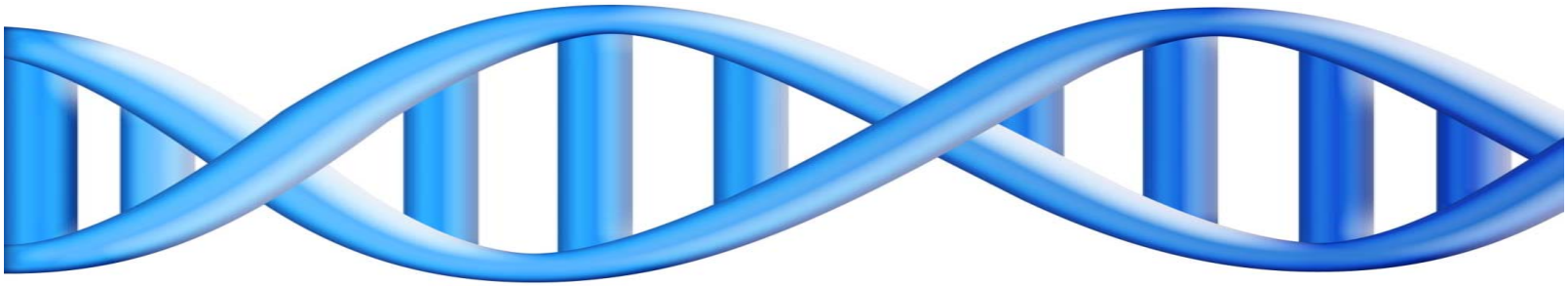


Integration and Demonstration Subprojects (ISP)

- Foster systematic evolution of project results
- Ensure horizontal integration of RTD contributions
- Result in integration prototypes

- ISPs are the main drivers during the project life-cycle

- They cause all work to converge (at least) 3 times in 30 months



OKKAM Infrastructure Evolution