ENGAGE: An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens

Nikos Houssos

National Documentation Centre (EKT) / NHRF
EuroCRIS

OpenAIRE Interoperability Workshop, University of Minho, 7/8 February 2013

Agenda

- ENGAGE project overview
- ENGAGE interoperability aspects
- ENGAGE collaboration opportunities

ENGAGE Project Information

Acronym ENGAGE

Title An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens

Website http://www.engage-project.eu

Platform http://www.engagedata.eu

e-infrastructure

Research Infrastructures

Contract no RI-283700

Project type CP-CSA

Start date 01/06/2011

Duration 36 months

Partners 9

SEVENTH FRAMEWORK PROGRAMME

Framework Programme 7 (2007-2013)

Project participants

NTUA (Coordinator)	GR
TU-DELFT	NL
MIC-GR	GR
IBM-ISRAEL	IL
INTRASOFT	LU
STFC	UK
FhG-FOKUS	DE
AEGEAN	GR
EUROCRIS	NL

Public Sector Information

- Data produced by governmental organisations – typically referring to datasets
- Examples: geospatial, demographic, statistical, environmental, public safety, financial data
- Growing international movement: open access to PSI datasets in a way that facilitates reuse
- Opening up PSI datasets can potentially lead to substantial economic gains [1]

[1] *Vickery, G. (2011): Review of recent studies on PSI re-use and related market developments.*

Overview of ENGAGE objectives

- Development and use of a data infrastructure, incorporating distributed and diverse public sector information (PSI) resources
- Capable of supporting scientific collaboration and research, particularly for the Social Science and Humanities (SSH) scientific communities,
- Empowering the deployment of open governmental data towards citizens.

Simply put, ENGAGE is a door for researchers that leads them to the world of Open Government Data. Through the ENGAGE platform, researchers and citizens will be able to search, browse, download, visualise and submit diverse and distributed Public Sector datasets from EU countries.

ENGAGE Two-way Scenario

•Public Sector Organisations •Open data initiations	•Pre-processing •Anonymisation •Harmonisation •Annotation •Linking	•Cloud and Grid Infrastructure •Platform Independence and Interoperability	•Open and intuitive access to the data collection •Context-specific search	•Visualisation (inc. combined views) •Context-specific formatting •Collaboration tools

Delivering Open Data Needs and guidelines to Public Sector Organisations

•Society •Policy	•Research Communities •Policy makers	•ENGAGE	•ENGAGE and eInfrastructures	•Public Sector Organisations

Ministries / local public agencies websites

Publicdata.eu

datos.gob.es

data.gov.uk BETA

Data.gov.be

data.gouv.fr

ckan

sdmx Statistical Data and Metadata eXchange

eurostat

National Statistical Offices

<ddi>

Unstructured / “Semi-structured”

Public data sources

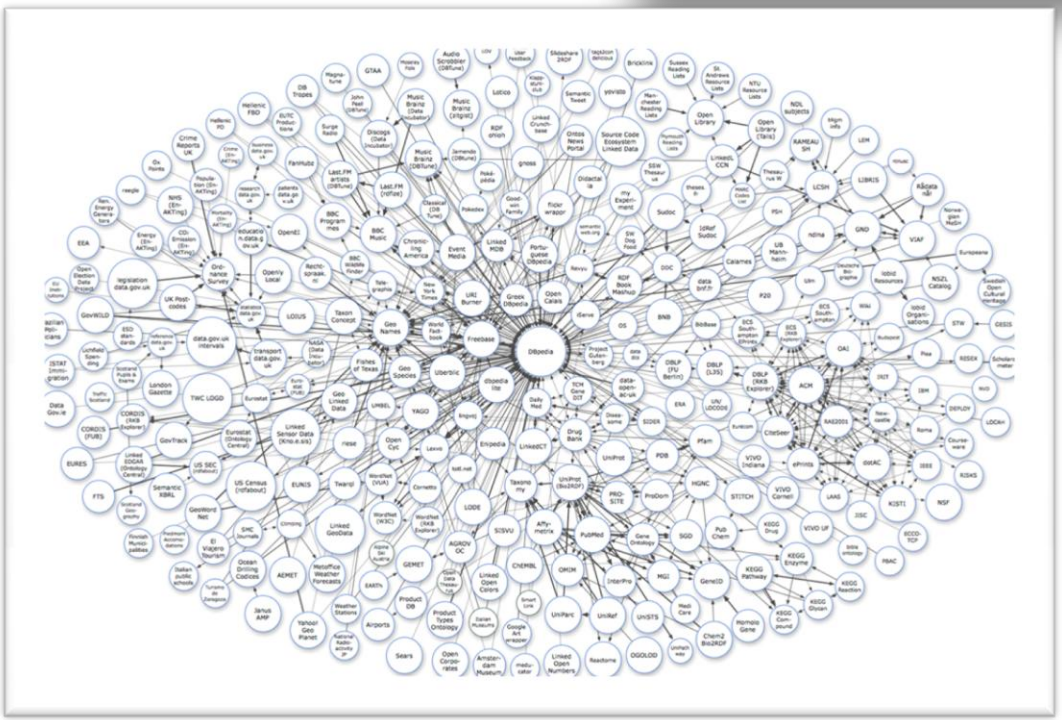
ENGAGE traverses across distributed and diverse public sector information resources

engage

API

ENGAGE provides a single point of access to PSI sources as well as relevant tools in order to cover the needs of researchers and citizens

ENGAGE: not an isolated data silo but a vital part of the Global Data Space.



ENGAGE aims to embrace the Linked Data Paradigm while ensuring the quality and responsiveness of highly structured information models.

ENGAGE will enable ***EU Researchers / Citizens*** to

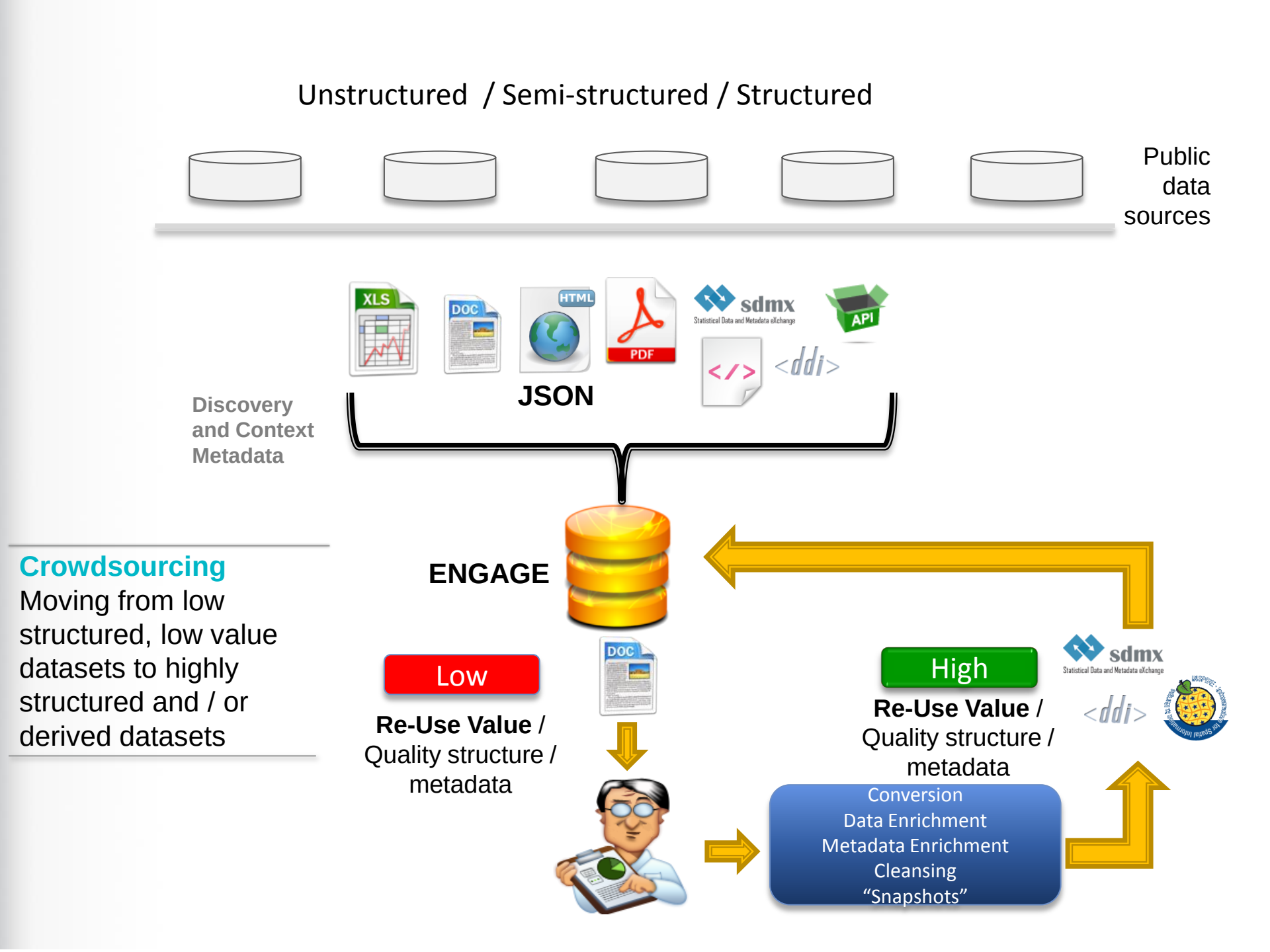
- ✓ Discover and browse datasets across diverse and dispersed public sector information resources (local, National and European) in their own language.
- ✓ Upload curated, enhanced or extended versions of existing datasets, originally published by public agencies, in order to address various formats, standards and scientific purposes in a *crowd-sourcing* manner.
- ✓ Acquire the datasets
- ✓ Visualize properly structured datasets in data tables, maps and charts

Additionally

- ✓ Utilize ENGAGE Application Programming Interfaces (APIs) for searching and acquiring the datasets.
- ✓ Rate the quality of datasets on various dimensions
- ✓ Request additional datasets or information on existing datasets from the Public Agencies
- ✓ View usage statistics
- ✓ View publications and other material linked to datasets

Public Agencies will be able to

- ✓ Utilize the ENGAGE infrastructure (interface and APIs) to publish governmental data
- ✓ Register and link their datasets within the ENGAGE infrastructure
- ✓ Receive feedback on the quality of their datasets
- ✓ Review the opinion or request of citizens and researchers
- ✓ View the applications, publications and other datasets uploaded by scientists, that are linked to their original published datasets


Unstructured / Semi-structured / Structured
Public data sources
XLS
DOC
HTML
PDF
sdmx
Statistical Data and Metadata eXchange
API
<ddi>
JSON
Discovery and Context Metadata
ENGAGE
Crowdsourcing
Moving from low structured, low value datasets to highly structured and / or derived datasets
Low
Re-Use Value / Quality structure / metadata
High
Re-Use Value / Quality structure / metadata
Conversion
Data Enrichment
Metadata Enrichment
Cleansing
"Snapshots"

engage

Natalia Kapelou
Political Sciences | Logout

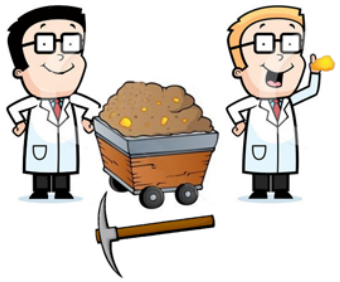
HOME DATA GUIDES / API WHAT'S NEW OPEN DATA SITES COMMUNITY CONTACT

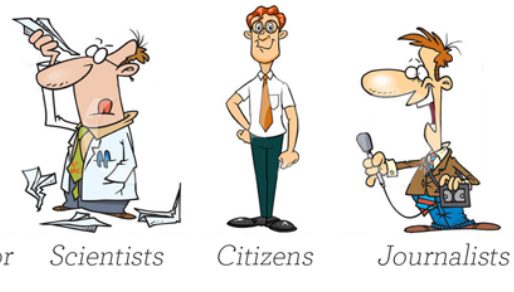
engage

Extracting value out of Public Sector Information

Learn More

Community-driven





for Scientists Citizens Journalists

Search Data

enter keywords...

+ Advanced Search

Request

I am looking for a dataset

- I request a format conversion
- I request metadata enrichment
- I request data cleansing
- I request data publication

Recent Submissions / Requests

Kartik Mandaville
Sociology

Kartik has enriched "*Criminal Incidents in Attica,Greece*" *in the DDI Format.*

Anna Morris
Political Sciences

Anna has performed data cleansing on "*Voting Results of France Elections 2012*".

Vicky Stamati
Political Sciences

Vicky has converted "*2011 Legal Decisions in the municipality of Tuente*" *from PDF to the Excel Format.*

Ministry of Finance
Greece

Ministry of Finance in Greece has published a snapshot of real-time data-set "*Greece Public Agencies Spendings*".

Top Research Karma

Joseph Mani
2134 Karma

Vicky Stamati
1260 Karma

Kartik Manda
812 Karma

George Doukas
112 Karma

Anna Moris
42 Karma

Top Public Providers

Ministry of Justice Greece
1720 High-Quality Datasets
2134 Datasets

Ministry of Justice The Netherlands
1105 High-Quality Datasets
1811 Datasets

London Gov Uk
341 High-Quality Datasets
702 Datasets

Israel Land Administration
65 High-Quality Datasets
432 Datasets

London School of Economics
36 High-Quality Datasets
102 Datasets

SITEMAP

HOME
DATA
GUIDES/API

LANGUAGES

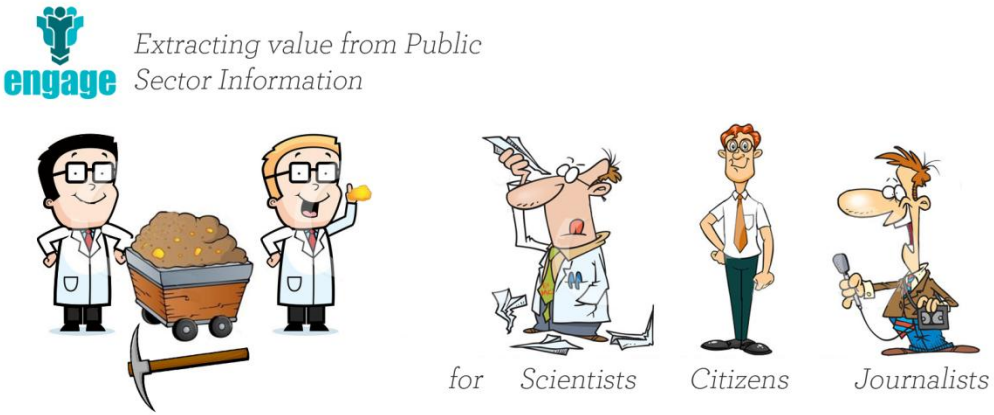
> English
> Deutsch
> español
> català
> Ελληνικά
> suomi

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

ENGAGE 2.0

- An infrastructure that integrates original PSI data and derived / curated datasets created, maintained and extended by users (researchers, citizens, journalists, computer specialists) in a collaborative environment. A curation platform with focus on the SSH research communities.
- To be released Spring 2013
- The vision of the ENGAGE infrastructure is to extract, highlight and enhance the RE-USE value of PSI data.
  - HOW: Moving from low-structured, isolated, difficult to find PSI data to easy to link , easy to process datasets with rich, structured metadata



ENGAGE 2.0

- On top of ENGAGE basic functions (catalog, search, visualizations, API)

Researchers / Citizens / Journalists:

- Extend other datasets (official or already extended - derived datasets)
  - Conversions (e.g. HTML- PDF to xls, PDF to RDF)
  - Data Cleansing (e.g. duplicate records, empty rows, errors)
  - Metadata Enrichment (missing metadata, Linked Data Enablers!)
  - Data Enrichment (enrich datasets with more information)
  - Snapshots of real-time data (e.g. Diavgeia_decisions_10_2012_to_12_2012.xls)
  - Mash-ups / Interlinking (e.g. Combine Election results to UV radiation levels!)
- View the version tree of official – derived datasets (clean solution - easy to understand and manage the contributions / versions)

ENGAGE 2.0

Researchers / Citizens / Journalists:

- Data Requests
  - Looking for a dataset (e.g. I can't find it elsewhere. Does it exist?)
  - Looking for a curation / conversion / enrichment (e.g. I am looking for the election results in Greece in XLS.)
  - Looking for data verification (e.g. Do you think this dataset is valid?)
  - Freedom of Information Requests

- Integration of tools
  - Google Refine
  - ScraperWiki
  - Visualizations

ENGAGE 2.0

Data Providers:

- Maintainers of Official Datasets
- Work as a group
- Bring the community which works on their data closer to them/ direct communication
- See and take advantage of ENGAGE Data Curation Community work (e.g. cleansing, better formats)
- Easy to see / gather all the Applications that are based on their official datasets.
- See the impact of their datasets.
- Understand which datasets have RE-USE value for users.
- Community Help in the process of Digitalization and Opening of current or older Public Data (history dimension)

Rich, structured metadata to enable Linked Data

- Structure: Entities and semantic relationships instead of plain fields
  - Each entity has structured metadata, including a URI field
- Semantics: Each relationship has clear semantics
  - What is the relationship of organisation Y with data set X?
  - Creator, maintainer, commissioner,…
- Ability to dynamically include into the system vocabularies => linked data, reuse of existing vocabularies/ontologies
- CERIF (Entities and Semantic Layer) provide the required features for contextual metadata

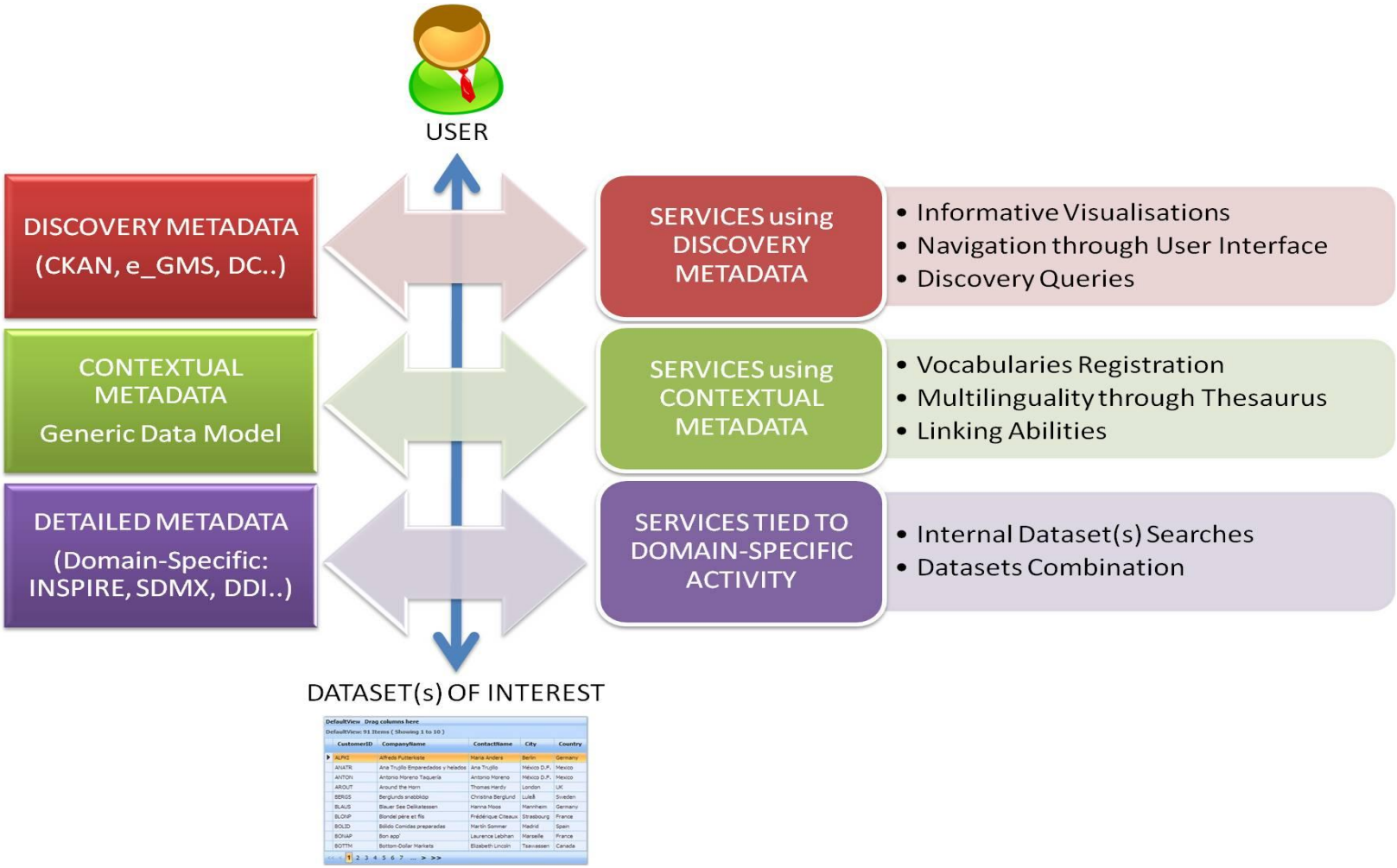
Rich contextual metadata is important

- Captures context, purpose, provenance, coverage, etc.
- Allows the user to:
  - Discover a dataset
  - Evaluate utility and re-use potential
  - Reuse it!
- Enables advanced services
  - Sophisticated search/discovery and navigation, mining, visualisation, reporting

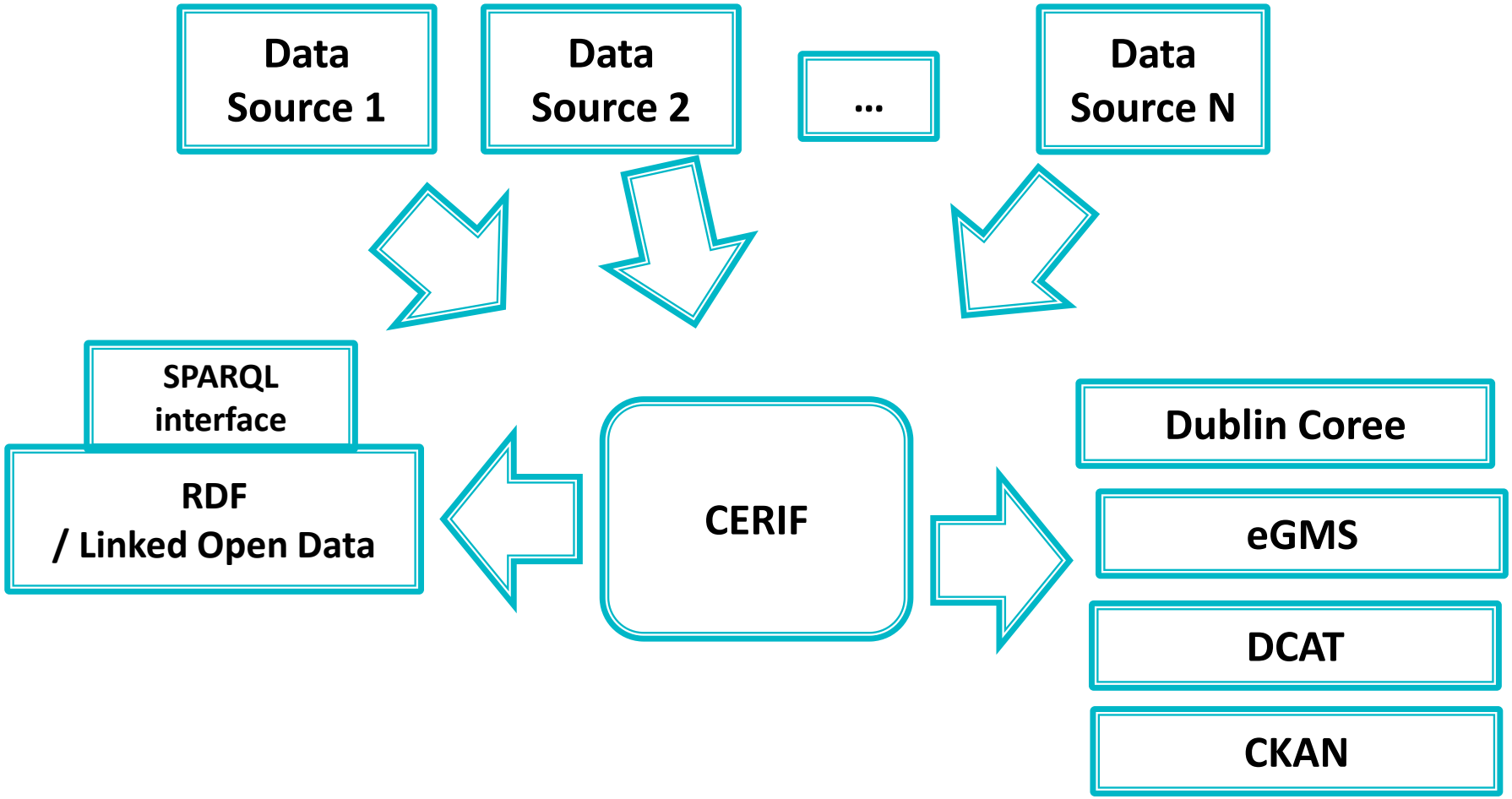
A 3-level metadata approach

- Level-1. Discovery metadata. Flat schemata (analogous to Dublin core). Enables basic search by non-sophisticated users.
- Level-2. Usage metadata. A structured, semantically-rich model for contextual metadata. Enables advanced domain-independent services.
- Level-3. Domain metadata. Detailed domain-specific metadata. Allows advanced services provided by specialised tools.

Metadata approach



Overview of architecture for PSI metadata



Thank you