



RISIS

Research infrastructure for research
and innovation policy studies



RISIS – Opening, access and interoperability of infrastructures

Research infrastructures for the assessment of science, technology and innovation policy

Matthias Weber, Thomas Scherngell and Peter van den Besselaar

European Commission, DG RTD, 01/14-01/17

Grant Agreement no: 313082

RISIS Annual Week, 26-29 January 2015, Rome

Agenda

- I. Overview
- II. Context for opening: Datasets
- III. Context for opening: Platforms

Overview

Matthias Weber

Why opening?

- Current situation

- Independent development of several datasets on STI aspects, based on customised approaches and definitions
- New opportunities arising from the availability of powerful IT equipment and big data approaches

- Expectations for the future

- Expanding use of meaningful data and indicators beyond standard I/O indicators
- Enabling synergies by combining different datasets in targeted projects
- Providing collaborative platforms to facilitate data analysis

➡ Opening of datasets as an opportunity to enhance the depth and quality of STI studies

Why harmonise?

- In order to exploit potential benefits of opening, there is a need to harmonise datasets in several regards to enable „inter-operability“:
 - Substantive harmonisation (i.e. shared categories, shared entities)
 - Access conditions (i.e. legal aspects, selection process)
 - Technical harmonisation (i.e. shared formats)
 - Common quality standards (i.e. transparency, reliability, stability)

Connectivity/substantive harmonisation

- Organisational entities
 - Register of organisations (WP 8)
- Geographical categorisation
 - Approaches to knowledge clustering (WP 9, kick-off)
- Economic and S&T categories (WP6)

Group Session B
27/1, 9-12:30

Group Meeting 3
27/1, 14-15:15

Group Meeting 10
28/1, 11-12:30

Harmonisation of organisational entities – register activities

- Developing a register of public research organizations (including research funding organizations and HEIs)
 - Introducing stable reference Ids
 - Tracking demographic events over time
- Dealing with organizational linkages and the presence of importance subunits
- Functions of the register
 - A reference list of organizations
 - IDs as a means to match together different databases and a solution to harmonization in the public sector
- Harmonization of private organizations to be dealt with as a second step.

Harmonisation of access conditions

Group Meeting 5
27/1, 15:45-17:15

- Specification of shared principles and conditions of access, as well as obligations for users
- Three main elements
 - Accreditation of users and compliance with basic RISIS principles
 - Selection process for research projects using RISIS datasets
 - Compliance with local access rules

Technical harmonization

Group Meeting 8
28/1, 9-10:30

Technical checks (WP 6.3)

- Writing a guideline with procedures and tools for
 - Data De-duplication
 - Data Disambiguation
 - Data Reconciliation
 - Data Consistency
 - Data Integrity
- Needed for individual datasets and for platform based integration

Semantic mapping (WP 6.3 & 7)

- Definition of entities
- Properties
- Scope, and related to that:
- Complementary datasets
- Vocabularies
- Approach: translation into RDF format

Common quality standards

Group Meeting 8
28/1, 9-10:30

- Transparent information about datasets and platforms
 - Standardised documentation about datasets and tools in order to facilitate access
- Shared quality standards for moving from experimental to robust datasets and tools
 - Documentation of updates/tracking changes
 - Tracing and documentation of use cases to enable learning
 - Stability and reliability

A joint pathway towards opening

- Checking and upgrading of datasets
 - Ensure fulfilment of substantive, legal and technical standards
- Accreditation infrastructure
 - Online facilities to apply for access
 - General RISIS accreditation + local accreditation
- Process
 - Step 1: Application for opening
 - Step 2: Making documentation and tools available (dataset reports, local conditions, etc.)
 - Step 3: Site visits and reporting (supported by checklist, trials)
 - Step 4: Official opening decision
- Making opening sustainable
 - Training courses to spread knowledge about use of datasets and tools
 - Collecting and learning from use cases
 - Collaboration with scientific networks (e.g. ENID)

Time plan for opening

Group Meeting 8
28/1, 9-10:30

<i>01/15</i>	Rome	Discussions on technical robustness/reliability framework, organisational/geographical categorisations, S&T/economic categories, accreditation and legal aspects
<i>03/15</i>	Online	Further exchange about framework for validation and categories to be used; agreement on framework and choices
<i>Start 04/15</i>	Application	Request for opening to FCB
	Site visits	Local site visits to check progress towards opening and reporting prior to opening decisions
<i>Start 04/15</i>	RISIS	Preparation of RISIS online interface for application and local preparations; project selection procedure
<i>Start 06/15</i>	FCB	FCB decisions to agree to opening of first individual datasets, based on results of site visit
<i>Start autumn 2015</i>	Facilities	Regular process of application for opening, including preparatory phase

Context for opening: Datasets

Thomas Scherngell

Main context for opening: 12 datasets

- Organisation of the opening of different RISIS datasets needs substantial preparatory activities concerning (WP6)
 - technical aspects, e.g. database system, data format, database location, potential interfaces (WP6)
 - substantive aspects related to database contents, e.g. harmonisation of sectorial and institutional classifications, denotations of organisations (WP8), geographical information (WP9)
 - legal aspects, e.g. access rights, IPR, obligations for users and providers

12 RISIS Datasets ...

... under consideration which can be attributed to five key topics:

- ERA Dynamics (3 datasets)
- Firm innovation dynamics (3 datasets)
- Public sector research in Europe (3 datasets)
- Research careers (2 datasets)
- Repository on policy evaluations (1 dataset)

Note: 10 Datasets already exist, 2 Datasets to be developed within RISIS

Datasets on ERA dynamics

- **JOREP (CNR)**
 - Comprehensive dataset on European trans-national joint cooperation programmes
 - in-depth insight into the FP research landscape in terms of thematic, organisational and geographical characteristics
- **EUPRO (AIT)**
 - Systematic information on research projects funded by the EU Framework Programmes (FPs) and all participating organisations
 - in-depth insight into the FP research landscape in terms of thematic, organisational and geographical characteristics
- **Nano S&T dynamics database (UPEMLV)**
 - Database containing publications and patents dealing with the production of nanotechnology over the time period 1991-2010
 - Information on about 1.18 million publications and 735000 priority, involving harmonisation of organisations and their location

Datasets on firm innovation dynamics

- **VICO (POLIMI)**
 - Dataset comprising information on new high-tech companies in seven European countries
 - Focus on the impact of venture capital (VC) on the economic performance of new high tech companies
- **CIB (UPEMLV)**
 - Information on more than 7.6 million patents (1986 and 2009) of 2000 large firms with the highest annual R&D investments
 - provides insight into technologies and geographical location of the corporate investments
- European mid-size fast growing firms (under development; **UPEMLV**)
 - A new experimental dataset on fast growing and mid-sized companies in Europe
 - Focus on geographical location, forms of innovation, etc.

Datasets on public sector research in Europe

- **ETER (USI)**
 - Register of Higher Education Institutions (HEIs) in Europe, providing data on various characteristics, such as
 - number of students, graduates, international doctorates, staff, fields of education, income and expenditure, among others
- **Leiden Ranking (UL)**
 - Information on the scientific performance of 750 major universities using a set of bibliometric indicators
 - measurements include the scientific impact of universities and of universities' involvement in scientific collaboration
- **European Public Research Organisation (under development, CSIC)**
 - Register of Public Research Organisation (PROs) in Europe, providing data on various characteristics, such as
 - number of staff, fields of science and technologies, income and expenditure, among others

Datasets on research careers

- **MORE (NIFU)**
 - Insights into researcher mobility patterns in Europe, the factors that shape them and the effects they can have
 - Based on a survey of EU-27 researchers (4538 valid responses), focusing specifically on their mobility events
- **ProFile (IFQ)**
 - Information on the situation of doctoral candidates at German universities and their postdoctoral professional careers
 - Based on a panel study (started in 2009) covering different topics, such as career intentions, scientific output, interruptions of candidacy, financing of doctoral studies, etc.

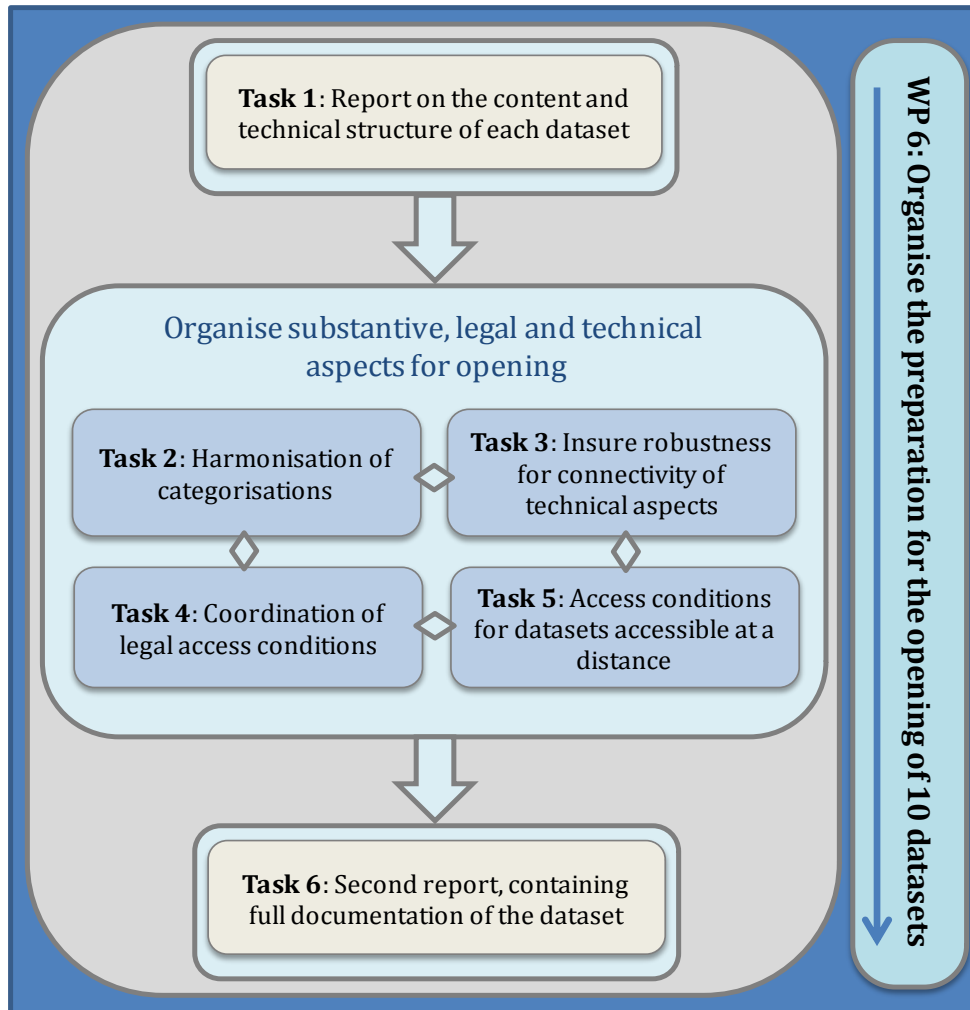
Repository on policy evaluations

- **SIPER (UNIMAN)**
 - database consisting of an on-line repository of about 1200 evaluation reports relating to innovation and science policy instruments
 - Enabling a structured search of information on the reports and their content, such as general topic, specific policy measure (based on a typology of policy measures), dissemination and quality issues, impact of the evaluation, among others

Properties and challenges of datasets

- mostly local use,
- very different technical choices which were selected for research developments rather than for favouring access and interconnectivity
- ad hoc solutions for solving disambiguation issues and for categorisations
- low levels of documentation (on their contents, but even more on their sphere of validity and on relevant conditions of use)

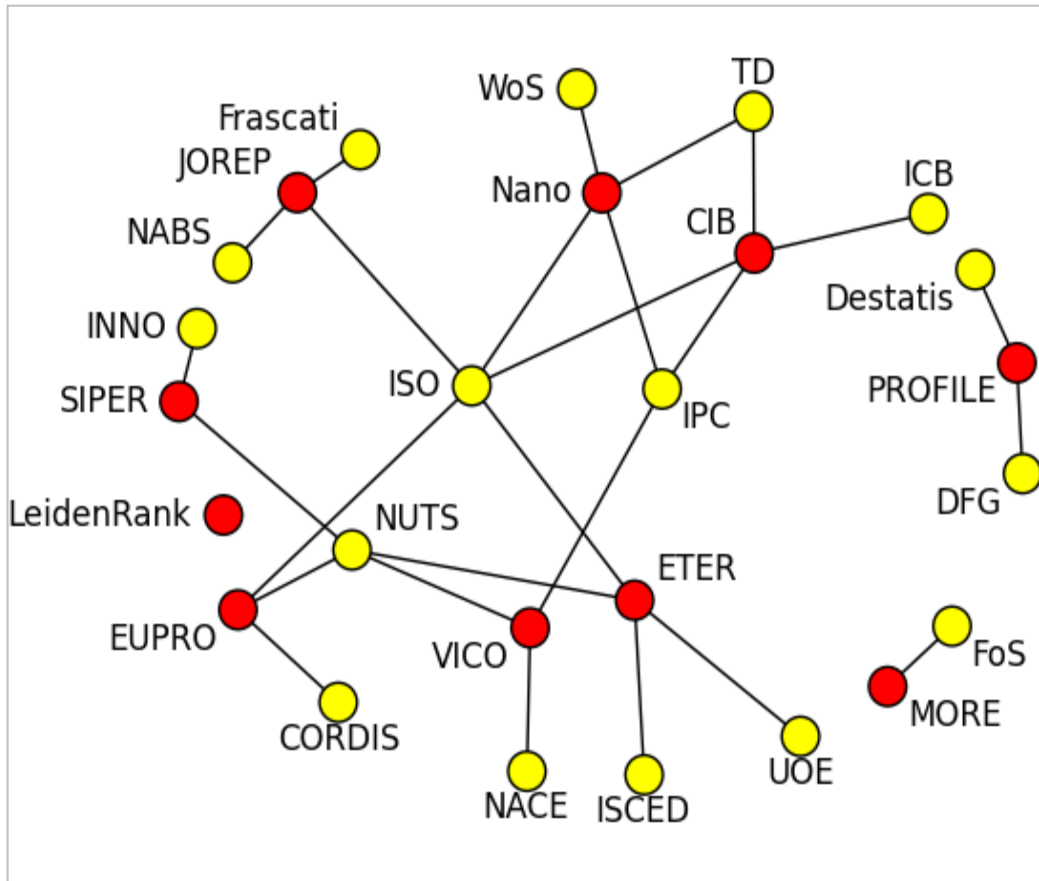
Workprogramm WP6



Status of Tasks

✓	Task 1	Establish a report on the content and technical structure of each datasets
<i>Started 10/14</i>	Task 2	Coordinate the work on the different categorisations used in the 10 datasets. A collective discussion will be initiated for all categorisations used in order to insure harmonisation
<i>Started 10/14</i>	Task 3	Stimulate an exchange with all databases holders to insure robustness and potential for connectivity of technical aspects
<i>Started 10/14</i>	Task 4	Coordinate work on principles and conditions of legal access of the datasets as well as obligations for users
<i>2015</i>	Task 5	Two datasets (SIPER, EUMIDA/ETER) that will be accessible at a distance require the access to be rebuilt
<i>2017</i>	Task 6	Produce a second report, containing a full documentation of the dataset taking into account the changes introduced through the collective work and explaining conditions of 'relevant' use

Overview on classifications used: A network perspective



CORDIS: CORDIS subject index

Destatis: Destatis Studienbereiche

DFG: DFG Subject Areas

EC: European Commission's list of associated / third countries in FP7

FoS: Field of science and technology classification (Eurostat)

Frascati: Frascati Manual (OECD, 2002) research topics / performing sector

ICB: Industry Classification Benchmark

INNO: Adapted INNO-Appraisal Scheme

IPC: International Patent Classification (WIPO)

IPO: Initial public offering

ISCED: International Standard Classification of Education

ISO: ISO Country Codes

NACE: Statistical Classification of Economic Activities (Eurostat)

NUTS: Nomenclature of Territorial Units for Statistics

TD: Technology fields of patents

UOE: UOE data collection on education systems

WoS: Web of Science subject classifications

A first exercise of joint exploitation of RISIS datasets ...

... in form of a combination of the **ETER** and **EUPRO** datasets

- Identification and analysis of determinants affecting participation of HEIs in the European Framework Programmes (see Lepori et al. 2014)
 - Combination of ETER and EUPRO using matching algorithms at AIT based on the organisation name
 - Creating a joint datasets combining HEIs participation (EUPRO) with HEIs characteristics (ETER)
 - Model the influence of these characteristics on participation intensity in a regression framework

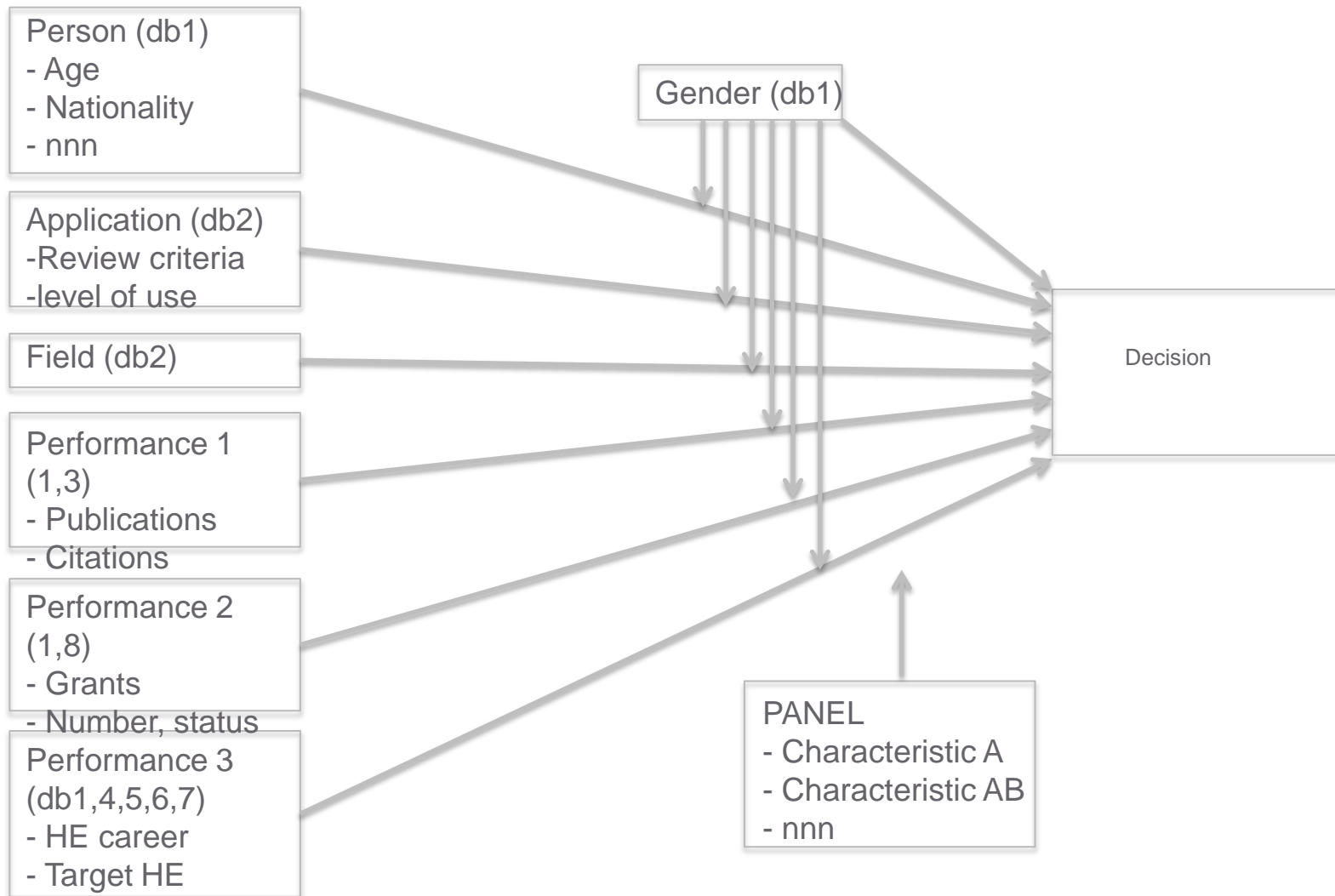
Context for opening: platforms

Peter van den Besselaar

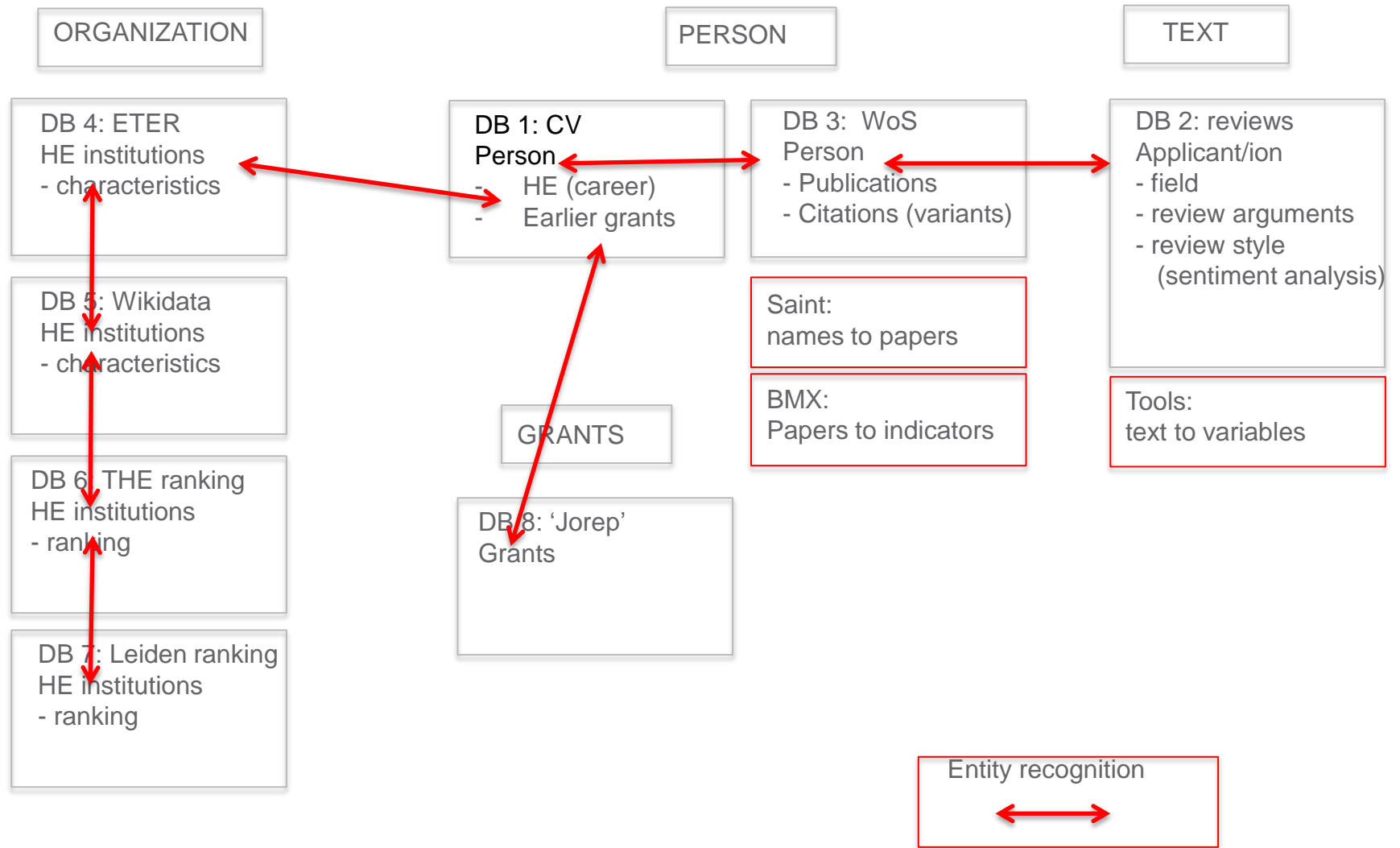
The platforms: CorTexT and SMS

- Aim: Distributed research infrastructure to support science and innovation studies
- Users: Researchers can use single or multiple datasets for advanced understanding analysis of the science and innovation system (and as *second aim*: indicators)
 - Accreditation and access
 - Local and global solutions
- How: Bridging (Data) Facilities and Platforms
 - Distributed datasets made interoperable
 - Specialized software platforms to support advancement of science & innovation studies

Example: gender bias in grant decisions

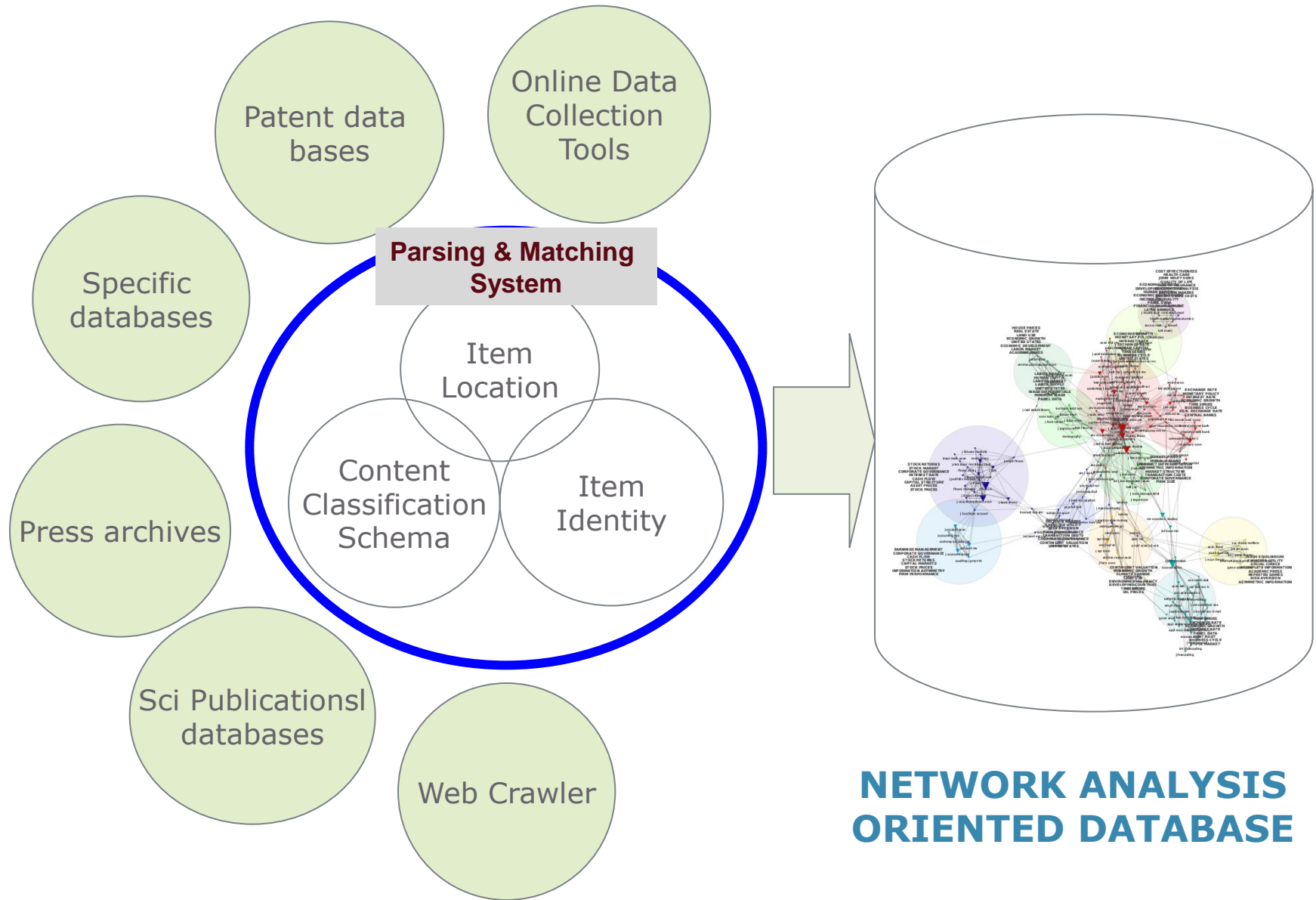


Data and tools

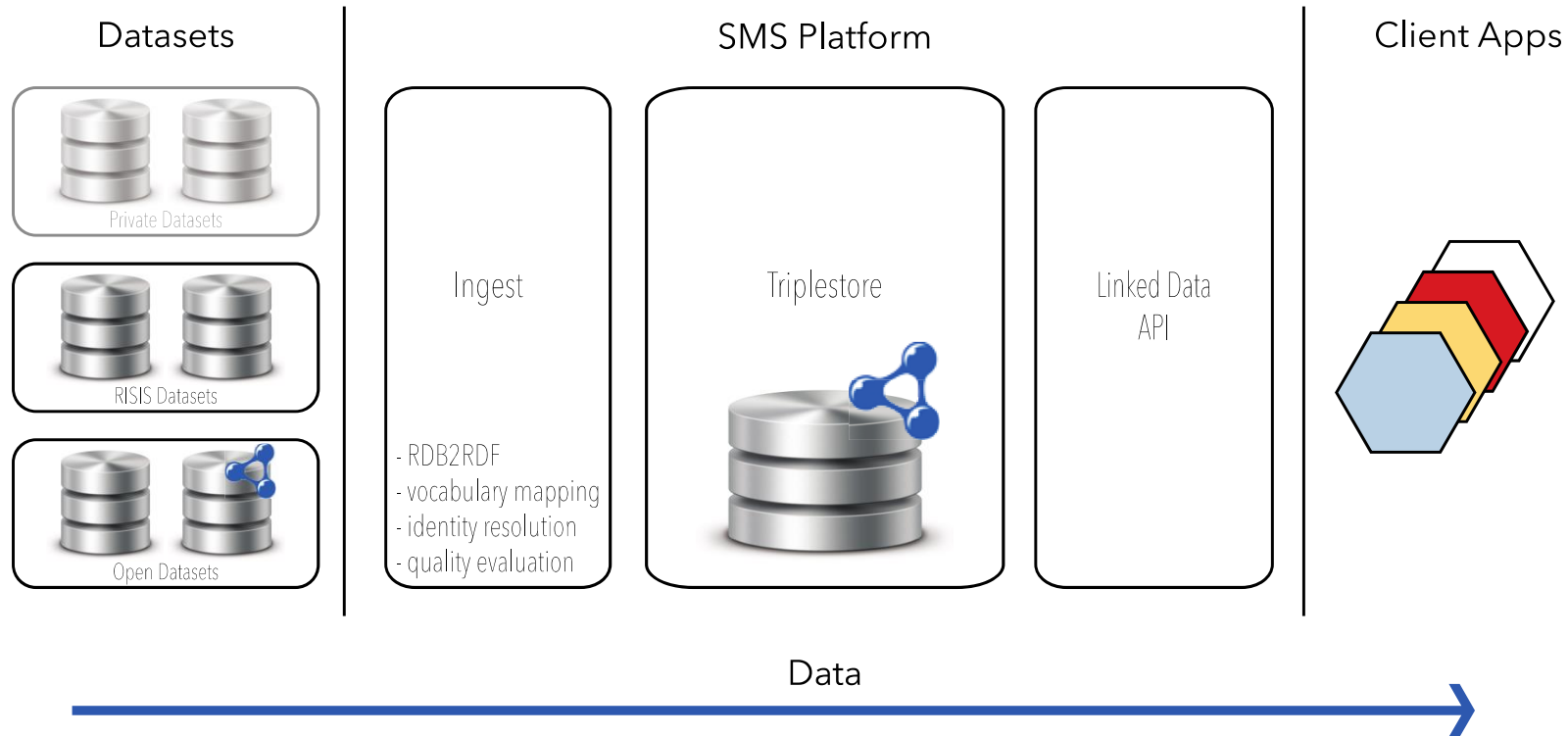


Dataset

CorTexT: complex back-office to access heterogeneous datasets for network analysis



SMS: Data integration with max flexibility, quality control, access



- Based on semantic Web Technologies for maximum flexibility and interoperability
 - Client access via a set of entity-centric APIs

Up to now: Risis-eu-CorTexT

- Developing parsers
- Parsing large set of datasets
- Technical infrastructure development
- Redesign core architecture
- Developing harmonized reference sets for interconnecting datasets

Up to now SMS

- Redesign of SMS approach using existing technology
- Testing the translation of several datasets in triple store format to enable integration
- Investigating the inclusion of citation/publication data
- Developing APIs (ongoing)
- First inventory of cleaning, processing, and disambiguation problems and solutions

Next steps (1) Developing the infra

- **Step 1:** Documentation about datasets + site visits to establish technical specifications for parsing solutions (CorTexT & SMS)
- **Step 2:** Designing a global parsing and integration strategy (CorTexT & SMS)
- **Step 3a:** Developing parsers for interoperability of Risis-eu CorTexT with each Dataset Platform (CorTexT)
- **Step 3b:** Further developing of data integration platform: integr. workflow, quality control, API, interface with tools (SMS)
- **Step 3c:** Testing the CorTexT-SMS interoperability
- **Step 4:** Technical tests and lead-users testing (CorTexT & SMS)
- **Step 5:** Release of a Beta Version when opening (CorTexT – M18; SMS – M26)

Next steps (2) : Making the infrastructure sustainable

- **Skills:** Training courses to spread knowledge about the use of analytical tools and data integration tools
- **Improvements:** Collecting examples of scholarly use case, as these inform about the needed data and analytical tools. This will inform tool and interface improvements
- **Relevance:** Collaboration with scientific groups and networks to specify the strategy of use cases in order to foster (joint) exploitation
- **Access:** Tools for access regulation, related to access rights

Platform sessions / Tuesday morning

Session A1 (9.00-1030h)

- Integration approaches
- Architectural issues
- Entity-Centric Data Integration (including disambiguation)
- Progress on CorTexT and SMS, and planning next steps

Session A2 (11.00-12.30h)

- Presentation CorTexT + discussion with users
- Presentation SMS + discussion with users “what are the important ‘standard datasets’ to get out of the system?”