COMMIT

WORKPLAN

WORKPACKAGES

DELIVERABLES

BUDGET

FROM DATA TO SEMANTICS FOR SCIENTIFIC DATA PUBLISHERS (P23)

Projectleader Prof.dr. Frank van Harmelen, Vrije Universiteit Amsterdam

1. Background

At the core of scientific development is the discovery of new knowledge; the generation, support and maintenance of knowledge form the foundation of the scientific endeavour. e-Science is ultimately about discovering and sharing knowledge in the form of experimental data, theory-rich vocabularies, publications and re-usable services that are meaningful to the working scientist.

The complexity and abundance of data resources in an e-Science environment requires support for knowledge and metadata management: data is notoriously hard to share, find, access, interpret and reuse. This project targets scientific data publishers as primary facilitators of the e-Science process.

Scientists need *tools* to better understand the complexity characteristics of their data and its ability to answer scientific questions. They must be able to equip data with *meaning* and to generate a surrounding semantic *context* in which data can be meaningfully interpreted. Scientists must be given the means to make their data speak for itself, to move from data to semantics.

The targets of this project are to

1. increase the ease with which scientists can *share* their datasets with others;
2. increase the ease with which scientists can *access*, *analyse* and *interpret* datasets, and thereby;
3. increase the *reuse* of such datasets.

To meet these targets, this project will develop a semantic infrastructure for data publishers, with facilities for finding, generating, tracing and interpreting scientific knowledge. It requires fundamental research on data complexity, knowledge acquisition, knowledge systems and services, and the development of a powerful set of tools that provides support for individual steps in the e-Science lifecycle.

We will collaborate closely with four COMMIT partner projects (P6, P12, P20 and P26), and develop against four use case domains:

- Data and publications of the COMMIT programme as a whole;
- Publishing workflow of Elsevier, one of the largest scientific publishers worldwide;
- Humanities research data provided by DANS, the Data Archiving and Networked Services division of the Royal Academy of Sciences in the Netherlands; and
- Health care research data used by Philips Research, Healthcare Information Management.

This is a revised version of the original P23 proposal that takes into account comments and advice from the IABC and COMMIT board.

2. Problem description

A core task for scientific publishers is to speed up scientific progress by improving the availability of scientific knowledge. This holds both for dissemination of results through traditional publications, as well as through the publication of scientific data. Data publishers face three key problems in fulfilling this task in an e-Science environment:

1. Distribution

   Scientific data is currently accessed and stored in isolated 'silos'. Individual researchers need to know exactly where to look for the data they need. Effective data sharing and reuse is only possible if data can be accessed transparently across trustworthy data repositories, unhindered by application specific restrictions.

2. Heterogeneity

   Scientific data is inherently heterogeneous; it is acquired, accessed and manipulated through different processes and methods, in different scientific contexts and at differing moments in time. Individual researchers need to know how to interpret the data they find. Enhanced access to data therefore requires that its provenance is made explicit, that its characteristics are understood, and that it is primed to be easily integrated and incorporated into software applications.

3. Provenance & Trust

   Traditional publishing methods rely on informal methods to determine the origin and trust of scientific datasets (personal communication, institutional reputation, social and scientific status, etc). In a world where locating, integrating and querying scientific datasets will be largely machine-mediated, we will need more formal and automated methods to determine the provenance and trust of scientific datasets that are available online.

In order to tackle these problems, we will focus on six challenges for data management in e-science: *how to share, publish, access, analyse, interpret and reuse data*?

How to share scientific data?

- Collaborative use and re-use of data requires the enrichment of data with knowledge from application scientists. This data enrichment can proceed through manual, community-based and automatic methods. It concerns identification of named entities, the recognition of patterns and regularities, and expressing such names and patterns as metadata using rich descriptive metadata vocabularies that are shared between scientists using open standards for both the syntactic form (XML) as well as for the metadata vocabularies (RDF, SKOS, OWL).[1]

[1] See http://www.w3.org/RDF, http://www.w3.org/TR/skos-reference, http://www.w3.org/TR/owl2-overview, respectively

- Can we enable researchers to work both in terms of commonly used (shared) concepts and in terms of personal concepts and hypotheses? How to manually, automatically, or semi-automatically construct or learn such vocabularies? This requires a strategy for the annotation of data with domain-specific ontologies, integrating local experimental data and community-wide knowledge.
- Can we provide the means to guarantee repeatability of e-experiments, provide high-quality provenance of data, manage data-lineage, and enable the answering of questions of trust and reputation? This requires metadata concerning provenance and lineage, suitable for building models and visualisation for trust and reputation.
- Can we bring these techniques and infrastructure within reach of individual researchers? Can we provide the tools to promote and support the management and traceability of knowledge all the way from initial data acquisition, through information, to structured knowledge?

How to publish scientific data?

- Data files constitute both the results of and the instruments for research. More and more journals require the open availability of data underlying the publications, dissertations refer to web sites, monographs contain digital attachments, etc. The difference between publications and data sets in the digital ecosystem is gradually fading. There are more and more 'enhanced publications' (in which sources, research results and data are taken from different web locations and then combined) as well as other forms, such as e-journals, with a data availability policy. Scholars and scientists need to have easy access to any information that is relevant to their research, whether it is data sets, publications, software (research tools) or project information.
- How can we make sure that the richness of data produced in an e-science infrastructure is not lost when research is disseminated through the traditional channel of scientific publications? In other words, how can we create, manage and maintain the links between (on-line) publications and data sets (and vice versa).
- How can we make sure that claims about scientific data in traditional publications can be individually referred to?

How to access scientific data?

- Access to scientific data requires querying over distributed and federated data archives. Engineering the required convertors from SPARQL queries to other formats (SQL, Lucene, etc) should be facilitated by easy-to-use transformation-generators, which must take local and domain specific conventions and semantics into account.

- Distributed reasoning concerns inferring implicit information that is not explicit in any of the local databases, but that can only be derived from their (logical) union.. Such distributed reasoning is crucial for semantic support in an e-Science environment where datasets cannot be physically joined. Current scalable approaches employ distributed forward chaining rule engines, which not scale to richer and larger datasets. Research in more economical backward reasoning approaches will allow tolerant, safe and scalable reasoning over uncertain and incomplete knowledge, embracing data, metadata and knowledge.
- Linking between vocabularies, establishing semantic connections between decoupled content, is a key element in transparent access to heterogeneous data sources. Both by allowing cross-discipline querying through vocabulary-based query translation, and by driving more traditional data integration. The field has made much progress in recent years, but is far from solved, with current recall and precision rates not surpassing the 80% range.
- In well curated collections, care is taken that each digital object has a single unique identifier. This assumption of 'a single unique identifier per object' no longer holds in heterogeneous and distributed collections, causing significant problems: can metadata attached to different identifiers for the same objects be combined? Given the combinatorial explosive nature of identity reasoning, can we still efficiently reason with such a proliferated set of identifiers on a large scale?
- Ranking is a key part in information retrieval but has received little or no attention in semantic queries. Can we identify meaningful and pluggable semantic distance measures that can be used for ranking results from query answering and search-engines over large e-Science corpora, and for establishing contextual relevance? Such measures can also be employed to tune search heuristics to priorities established for a given domain or researcher, from query building and auto-completion functionality to session persistence.

How to analyse scientific data?

- Is the difference in methodology between different disciplines reflected in the ontological structure of their domains? For instance, science and the humanities have a strikingly different structure. Humanities predominantly deal with large classes of inherently different individuals, isolated events and structures (a painting by Van Gogh, the battle of Waterloo). Empirical sciences predominantly deal with large classes of highly structured publicly accessible events or object (e.g. atoms, diseases). This difference has large consequences for the structure and the accessibility of the data sets that are to be created in these domains.

- How can we deal with the complexity of scientific models? An adequate model of e.g. the human brain might have a complexity in the order of 10^15 bits. This is orders magnitude more complex than the systems and models that empirical sciences have studied up till now. Not only will theory formation demand the creation of extremely large data-sets that cannot adequately be overviewed by any individual (or even group of scientists), it is also the case that from this complexity point of view our universe is fundamentally undersampled. There are currently 'only' in the order 10^9 human beings. The 'traditional' empirical approach of first gathering an overload of data and then compress these data in to a theory, will not work in these domains. New methodologies, e.g. model-based reasoning, will be needed to crack these domains.
- How can we assess whether a particular dataset is informative enough to be suitable for answering a query? This relates to complexity classes: recent research has come up with various new definitions of meaningful information, e.g. Facticity (Adriaans 2009). These developments allow us to analyse and organize the various issues in scientific data management from a new perspective.
- How can we employ these insights in information utility for guiding and driving distributed access to data sources?

How to interpret and reuse scientific data?

How can we facilitate researchers in interpreting data that has been shared, published, accessed and analysed in the previous phases? This requires the development of methods and tools for visualising and interpreting data and context in multiple dimensions such as time, place, trust and provenance, scientific field, etc.

- How can we encourage that the results of data interpretation are fed back into the pool of annotations surrounding that data?

3. Objectives

Project's goal

The availability of hitherto unimaginable volumes of scientific data forces us to rethink the scientific methodology and our modes of gathering and organizing data for science. At the core of scientific development is the discovery of new knowledge. Scientists must be empowered to better understand the complexity characteristics of their data and its ability to answer scientific questions. They must be able to equip data with meaning and to generate a surrounding semantic context in which data can be meaningfully interpreted. The goal of this project is to

1. substantially increase the ease with which scientists can share their datasets with others;
2. substantially increase the ease with which scientists can access, analyze and interpret datasets in their domain of inquiry, and;
3. to substantially increase the re-usability of such datasets.

Scientists must be given the means to make their data speak for itself, to move from data to meaning.

Planning of all dimensions

The goal of this project is to

Goal 1. increase the ease with which scientists can share their datasets with others,

Goal 2. increase the ease with which scientists can access, analyse and interpret, and

Goal 3. increase the re-usability of such datasets.

These goals translate into the following concrete scientific and technical objectives:

Obj 1. A theory on the suitability of a dataset in answering a query, based on complexity measures (WP1).

Obj 2. Methods & tools for cross-vocabulary querying of distributed datasets (WP2).

Obj 3. Methods & tools for ranking query-results based on semantic distance measures (WP3).

Obj 4. Methods & tools for publishing and integrating linked data (WP4).

Obj 5. Methods & tools for tracking, representing and visualizing provenance of scientific data (WP5).

Obj 6. Methods & tools for annotating and visualising scientific data (WP6).

Achievement of these goals will be demonstrated in use-cases:

Case 1. Elsevier: Assertion-level Document Annotations in Pharmacology.

Case 2. DANS: Linked Data for e-Humanities.

Case 3. Philips: Linked Data for clinical decision support in prostate cancer (to be decided).

Case 4. Publication of a COMMIT Linked Data set

To meet these targets, we have distributed work over six work packages, that each target specific aspects of the e-Science workflow (see the picture below). WP1 and WP2 will develop a theory on the suitability of a dataset in answering a query, based on complexity measures, and the implementation of tools that exploit this theory in building queries and answering them in a distributed fashion. WP3 and WP4 aim to improve how scientific data is shared and accessed, by developing new distance measures for ranking query results, and tools for translating, interlinking, enriching and reasoning over scientific data. WP5 and WP6 are concerned with traceability of scientific results. In WP5 by analysing how data and data provenance is to be used by researchers and represented as linked data, and in WP6 how this data is to be presented, visualised and improved.

In terms of the 5 COMMIT dimensions, the planned results are as follows:

- Results through methods as listed in Obj 2-6, software tools as listed in Obj 2-6, every WP is committed to publications in high impact journals and conferences
- Impact through product transfer of the use-case results to Elsevier, DANS and Philips

- Dissemination through inclusion of material in courses, summer schools, social media (specific plans to follow as the project progresses)
- International through connections with key EU projects and through standardization activity in W3C.
- Synergy through well specified collaborations with
  - (we use their technology),
  - P12 (we collaborate on provenance technology)
  - P20 (they use our technology),

Results

The main types of Results from P23 will be a Golden Demo, software prototypes, datasets and publications:

- The Golden Demo concerns the integration of Linked Data in scientific publications in the domain of Pharmacology, showing on a publication corpus from Elsevier the benefit for researchers. The functionality of this Demo concerns the simultaneous navigation of a researcher through publications and research data, while tracking provenance of both.
- Software prototypes will be produced by each WP at every annual milestone. Many of the software prototypes are expected to feed into the Golden Demo.
- E-Humanities datasets will be produced based on data available at DANS. A COMMIT Linked Dataset will be produced, containing both scientific data and project meta-data (publications, presentations, social network, etc). We will also help other projects to publish their data as Linked Data (arrangement with P26 has been made).

Deliverable Impact and Valorization

In the first year, we expect the impact of P23 to concentrate on knowledge-transfer to our non-academic partners. DANS: A representative of the technical partners will work part-time in house at DANS, ensuring an optimal environment for knowledge transfer. Elsevier: The year-1 demo) (linking data and publications) will be closely tied to the primary publication process in Elsevier.

By year 2 we will use data from other COMMIT partners (P6, P20, P26) to further test and develop our methodologies, algorithms and tools. We will thereby showcase the technology to the non-knowledge partners in these projects, allowing for a wider adoption of our results. P6 and P20 play an important role both in data and technology exchange. Furthermore, our contacts with historians in WP5 will bring our technology to researchers in the humanities. Outside the COMMIT consortium, we expect significant interest both from the pharmaceutical industry, as well as the non-profit and government sectors for the technology developed in P23: in all of these sectors, awareness of the value of Linked Data technology is rapidly growing, but adoption of this technology is currently hindered by a high threshold. To foster transfer to the pharmaceutical sector, we will exploit our strong links with the EU-funded IMI project Open-Phacts (Open Pharmacological Space). The final channel to ensure impact is our participation in a number of W3C standardization working groups concerning the technology of P23: linked data technology (SPARQL WG, RDF WG), linked data for health care (HCLS WG) and provenance (Provenance WG).

Deliverable Dissemination

Because P23 is a project about infrastructure, our dissemination will be focused on a professional, using the following tools and channels:

- Project website to be established in first month,
- All papers, reports and software downloadable from the website
- Exploitation of social media (project leader and envisaged postdocs are already active)
- Usage of SlideShare and Mendeley (including Linked Data feeds)
- Usage of CKAN for registering datasets
- Usage of Sourceforge or Google Code (to be decided) when allowed by Project Agreement
- Every demonstrator will come with a screencast

Although the general public is out of scope for this project, work package leaders have excellent contacts both with national media (written press as well as radio), evidenced by numerous interviews over the preceding years.

International Imbedding

The central idea of P23 (lowering to share, access and re-use datasets on the web) is well embedded in an international context. P23 is also unique in a number of respects:

- The participation of Elsevier, one of the largest scientific publishers, gives us access to real-life data and use-cases, and enables P23 to directly influence industrial practice.
- The participation of DANS as one of the largest archives of social science data in Europe enables P23 to step across the gap between the natural sciences and the social sciences.
- Although many other projects aim at technology and infrastructure for scientific linked open data, P23 really aims to lower the threshold for non-computer scientists.

Deliverable Synergy

P23 expects to receive deliverables from P6, P12, P20 and P26 in M6 of Y1 (at the start of Q2). These deliverables will encompass requirements and use cases for (linked) data usage in the respective domains of these projects: cultural heritage and trust, public safety, concept-centric networks, and food science. In turn, we will deliver the showcases developed on the basis of these and P23- internal requirements at the end of Y1 (Q4). P6 will also provide requirements for the provenance representation in WP5, this will be a joint deliverable in M6. We furthermore expect a deliverable on large-scale reasoning from P20 at the end of Y1 (Q4) to be integrated with tools developed in WP4. Year 2 will bring stronger ties with COMMIT partners as we will start using their technology (from P20 and P6) and their data (from P6, P20 and P26) to further develop P23 technology. These results will be delivered back by the end of Y2, as well as technology transfer to P6 (provenance) and P26 (agrifood). Years 3 and 4 will offer further refinement based on continuous evaluation in close contact with COMMIT synergy partners.

4. Economic and social relevance

The economic and social relevance of P23 is ensured through the participation of three key stakeholders in both aspects of the e-Science lifecycle: Elsevier Publishing, the Data Archiving and Networked Services (DANS) institute, and Philips Healthcare Information Management.

The production of scientific knowledge underlies most if not all of our sustained growth in wealth. Improving the efficiency of the "scientific cycle" is therefore crucial to modern economies. This project will contribute to increasing this efficiency by reducing the "cost per discovery", reducing the "time per discovery" and increasing the "quality of discovery".

Reducing costs of scientific results: In many scientific areas, datasets that have been produced at high cost in equipment and labour are locked up in individual labs. Reuse of datasets between scientists across the globe is often limited due to restrictions in the available infrastructure: data-formats differ between labs, terminologies differ between subfields, provenance information is missing, and data is often not available at the right time at the right place. Removing such restrictions would result in much higher re-use of such expensively produced scientific data.

Speeding up the scientific cycle: not only is data production (also known as "measurements", "observation", etc.) the most expensive part of the scientific cycle, it is also be the most time-consuming. Scientific discovery could be greatly accelerated if data-sets would not have to be re-produced at different times at different locations by different organisations, but if instead existing data-sets could be located, accessed, transformed, integrated and deployed at low cost in both man-power and infrastructure.

Increasing scientific reliability: although re-production of scientific results and the repeatability of scientific experiments are at the methodological heart of the scientific process, the actual repeatability of scientific results (and the corresponding confirmation or potential refutation of results claimed by others) is surprisingly low in the daily life of a working scientist. The reliability of scientific results would increase substantially if methods and technology are available for re-using each other's datasets, and for tracing the lineage of data used in experiments performed by others.

Generic results that are important to other economic sectors: in our information-based economy, sharing of data is important not only to many economic sectors:

- With an increased pressure on the pharmaceutical industry to develop new drugs, there is an apparently corresponding shift toward data sharing occurring. The pharmaceutical industry is starting to embrace semantic technologies and linked data. While the adoption of linked data is still not yet very widespread in individual companies, it is on the agenda of several large-scale cross-pharma projects such as OpenPhacts, the Pistoia Alliance and the BioRDF taskforce of the W3C HCLS interest group.
- e-Commerce and e-procurement are economic activities that are of increasing importance and which also crucially rely on the ease, reliability and speed with which data can be shared and trusted between different parties (in that case between different parties in a particular value-chain).
- The cultural sector is increasingly moving towards sharing of data, e.g. with virtual collections that span multiple museums, cross-collection search capabilities among archives, etc. Again, such developments crucially rely on the ease and reliability with which data can be shared between different parties. Such innovative cultural activities can be expected to have indirect economic value through spin-off in for example the tourism sector.

5. Consortium

The scientific core of the consortium is formed by three technical partners at VUA and UvA:

- Informatics Institute, Universiteit van Amsterdam (UvA) Prof. Pieter Adriaans, lead partner for WP1 and WP2;
- Knowledge Representation & Reasoning group, Computer Science, Vrije Universiteit Amsterdam (VUA) Prof. Frank van Harmelen, project coordinator and lead partner for WP3 and WP4;
- Web & Media group, Computer Science, Vrije Universiteit Amsterdam (VUA) Prof. Guus Schreiber, lead partner for WP5 and WP6;

Three leading stakeholders in the e-Science lifecycle will participate as non-academic partners. They will ensure the economic and social relevance of project results, and will provide use cases and requirements:

- Elsevier Publishing (ELS), world leading scientific publisher, and
- Data Archiving and Networked Services (DANS) institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO), responsible for the permanent storage and accessibility of research data in the arts and humanities and social sciences.
- Philips Research, Dept. Of Healthcare Information Management.

The P23 consortium will collaborate with at least four other COMMIT projects:

- P6 (Socially-Enriched Access to Linked Cultural Media),
- P21 (Dependable Cooperative Systems for Public Safety),
- P20 (e-Infrastructure Virtualization for e-Science Applications), and
- P26 (eFoodLab – Disclosing data and knowledge in food research).

This collaboration covers:

- datasets made available from P6 and P26 to this project and enriched web-enabled datasets being returned to those projects.
- technology for large-scale reasoning developed in P20 and deployed in this project in WP4.
- provenance infrastructure developed in WP4 and WP5 will be used in P6 and P12.
- data querying and publishing technology developed in WP2 and WP4 will be deployed and evaluated in P20 and P26.

Details about these collaborations, and the synergy deliverables exchanged between the projects are specified in the yearly plans. Internationally, we will closely collaborate with a number of EU projects (LATC, LOD2, PlanetData, OpenPhacs, DARIAH, EU-ADR). We refer to the section on distribution and transfer of knowledge for more details.

Elsevier (ELS)

As the world's leading publisher of science and health information, Elsevier serves more than 30 million scientists, students, and health and information professionals worldwide. Elsevier's mission is to contribute to the progress and application of science, by delivering superior information products and tools that build insights and enable advancement in research. Elsevier products and services include 1,800 journals and over 50,000 books, reference works, and textbooks written and edited by international scholars with outstanding professional credentials in their field. The project will develop showcases that will help Elsevier find new directions and demonstrate future Elsevier capabilities on already existing data, resulting in four potential showcases:

1. Using RDFa in Elsevier content. Elsevier aims to include RDFa representations in its publications to enable improved data integration, both within Elsevier publications and between Elsevier and other publications. This showcase will be driven by requirements from both Elsevier and DANS.

2. Paragraph-level identifiers to enable fine-grained links to paragraphs in Elsevier publications from external sources. This showcase will be driven by requirements from Elsevier and DANS.
3. Alignment of Elsevier semantic content with the linked data cloud. By adding reciprocal links from Elsevier content to external linked-data sources, Elsevier semantic content can be placed into context.
4. Supporting the annotation process at authoring time. This is a very important enabler for enriched publications, and it allows authors to provide meta-information on the level of certainty of their claims and the links between claims and evidence.

Elsevier provides a strong business case and real world requirements for the linked data e-science infrastructure developed in P23. Elsevier will play an active role in the design, development and testing of showcase scenarios in WP4 and WP6. Access to Elsevier content and infrastructure will provide a unique test bed for linked data use in e-science, allowing P23 partners to research and develop their technology with real data, and demonstrate its socio-economic value.

Data Archiving and Networked Services (DANS)

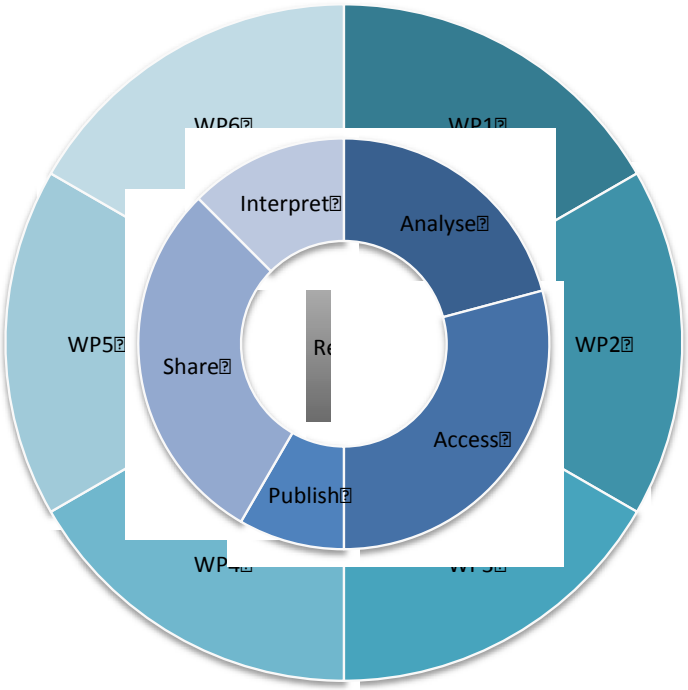
Research data in the social sciences and humanities are often heterogeneous and complex in form and content. This has to do both with levels of standardization across scientific disciplines, but also with the highly complex nature of social systems and cultural heritage as objects of study. Data reuse is only possible if the methods for collecting the data and its provenance, are precisely documented and communicated. The time and location at which data is collected are crucial for determining its value, noteworthiness and role in research. These specific requirements bring fundamental challenges for data management. DANS will provide access to a variety of different data collections across the social sciences and humanities, and has excellent connections to the corresponding knowledge domains for both data and expertise. Furthermore, like Elsevier, DANS brings extensive knowhow and practical experience with permanent storage and data accessibility (data management, metadata standards, quality guidelines, data migration, persistent identifiers, enhanced publications, licensing, etc.).

Through participation in P23, DANS aims to disclose their data collections through an advanced linked data infrastructure, integrate them with other datasets, and explore the added value of this technology for researchers in the social sciences and humanities. Secondary goals are the development of visual aids for researchers in data provenance and collection browsing. This will result in three potential use-cases:

1. Publishing the Dutch national census data from 1795 to present as Linked Open Data
2. Publishing the E-Depot of Netherlands Archaeology (EDNA) as Linked Open Data
3. Enhancing scientific publications with meta-data from the Dutch Research Database (NOD)

6. Work plan

The P23 general work plan uses a spiral model. Each cycle will deliver a significant result,



indicated by major milestones. While work packages can still customize the timing to address their specific needs, general milestones are established for three phases at every iteration:

1. analysis phase: Key outcomes are research papers, requirements, use-case scenario's, initial
2. designs of semantic models, exploitation and dissemination plans;
3. development phase: Key outcomes are prototypes and showcases;
4. evaluation phase: Key outcomes are end-user evaluation, deployment in other COMMIT projects. The outcome of each iteration is evaluated and forms the input for the analysis phase of the next iteration.

This project is structured in 6 work packages that correspond to the research questions. We have taken care to balance external dependencies for work packages that involve PhD work. Each work package targets specific aspects of the e-Science workflow.

We briefly describe each work package below, full work package descriptions are included in section **Error! Reference source not found.**.

WP1.	Information content and utility
This work package focuses on the analysis of data, and will develop a theory on the suitability of a dataset in answering a query, based on complexity measures.	
Leader	Pieter Adriaans

Staff	PhD + UD/Prof
Synergy	WP2
WP2.	Support for querying distributed complex data sets
This work package focuses on data access, and will implement a distributed query platform. It exploits the theory of WP1 in determining what resource to query, uses the interlinked vocabularies of WP4 to enable transparent cross-vocabulary querying, and provides a platform for query results ranking developed in WP3.	
Leader	Pieter Adriaans
Staff	SE + UD
Synergy	WP3, WP4, P20
WP3.	Distance measures and ranking of query results
This work package aims to improve data access and will design, develop and evaluate new distance measures for ranking query results. It will use and feed into the query platform of WP2, and relies on the enriched datasets produced in WP4 as well as use cases for evaluating ranking strategies.	
Leader	Frank van Harmelen
Staff	PhD + UD
Synergy	WP4, WP2[2]
WP4.	Information publication, integration and enhancement
This work package focuses on data sharing and publication, and will design, develop and evaluate techniques for publishing, integrating, enriching and reasoning over linked science data. These are closely tied to requirements of use case partners. It will use the data provenance representation and annotations of WP5, as well as align representations with requirements from use case partners and WP2 and WP3.	
Leader	Frank van Harmelen
Staff	PD + SE + UD/Prof
Synergy	WP2, WP3, WP5, P20[3], P6, P26
WP5.	Information provenance
This work package focuses on data sharing, and will design, develop and evaluate methods and tools for scientific data provenance, covering representing, tracking, annotating and visualizing provenance data. How are data and data provenance to be used by researchers and represented as linked data? Work will proceed in close collaboration with WP6 and P6 and P12. Results will feed into WP4 and WP5 as well as P6 and P12.	
Leader	Guus Schreiber
Staff	PhD + UD
Synergy	WP6, P6, P12
WP6.	Information annotation and interpretation
This work package focuses on data sharing and interpretation, and will design, develop and evaluate methods	

[2] Results can be evaluated independently from use in **WP2**.

[3] **P20** will provide large scale reasoning technology to be used in data enhancement.

and tools for annotation, interpretation and visualization of scientific data. How is data to be presented, visualized and improved?	
Leader	Guus Schreiber
Staff	PD + SE + UD/Prof
Synergy	WP2, WP4, WP5, P6, P20, P26

The services developed in all work packages will be introduced as components for the workbench developed in WP6, including components from other COMMIT projects, such as P6 for analysis and visualisation of trust, and P20 for large scale reasoning. We will adopt a plugin architecture to reduce the risk of overly depending on other work package results. Although every component can be evaluated independently with respect to the challenges of section **Error! Reference source not found.**, the workbench will allow us to move beyond the sum of all parts by forming a full e-Science data-usage infrastructure.

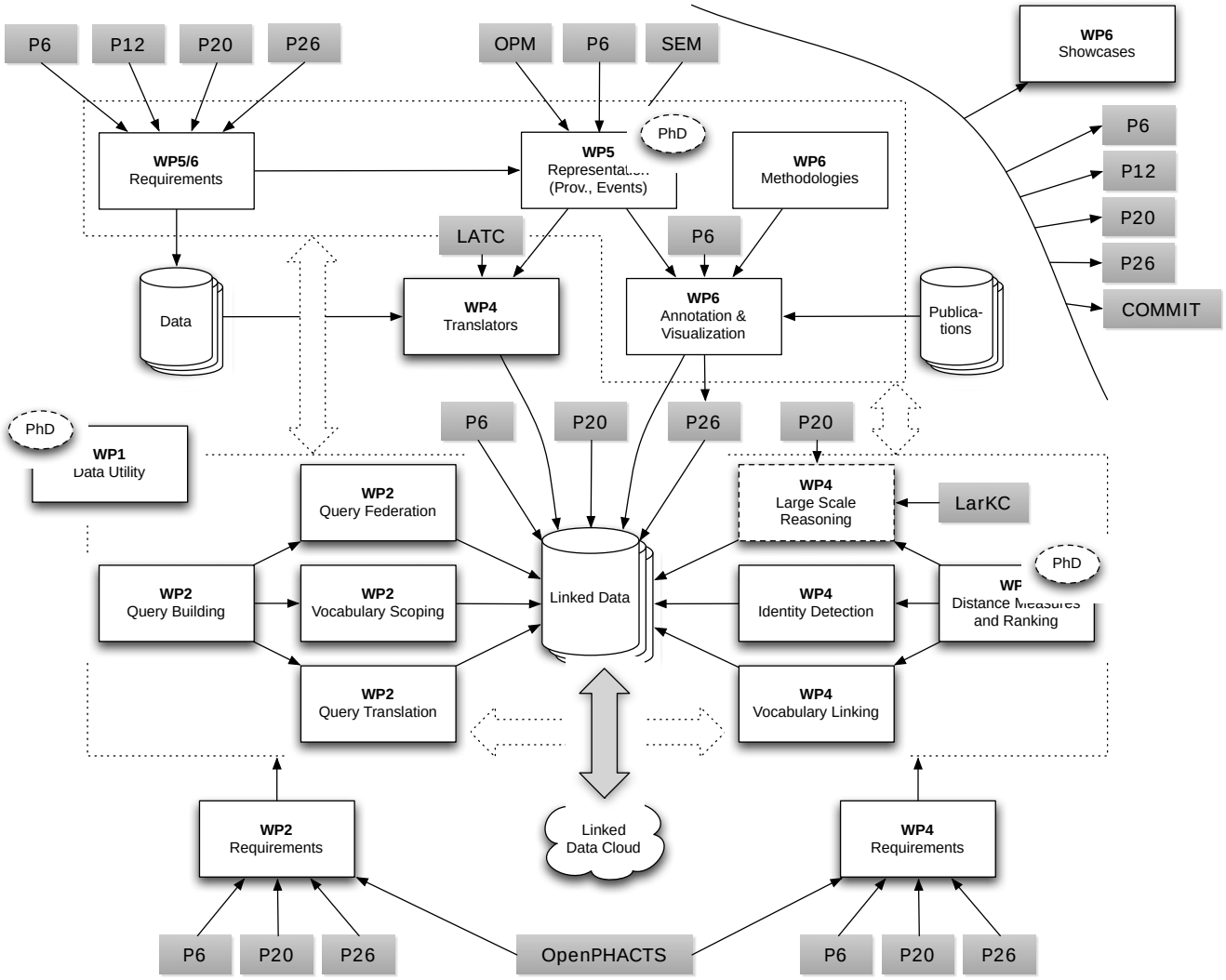


Figure 2: Integrated schema of dependencies between work packages and other projects

Elsevier Use Case (Golden Demo)

To illustrate the work plan and connections between work packages, this section describes a typical use case as will be addressed by the P23 project. We describe the use case from the perspective of an individual researcher. It illustrates the type of functionality scientific data publishers seek to offer their users. Assertion-level Document Annotations in Pharmacology

Motivation

Pharmaceutical researchers need to be able to have access to an integrated view of all of their data in order to be able to make effective decisions as to which drug targets and compounds to pursue. Companies want to minimize costly late-stage attrition by identifying and eliminating drugs that do not have desirable safety profiles or sufficient efficacy as early on as possible. The need for effective data integration has become stronger as the cost of drug discovery and development has soared to over $1 billion.

Scenario

A scientist working on (drug) target optimization tries to understand the relationship between a gene and the expressed protein. She would like to know everything that is known about a given gene, and more particularly the effect of an engineered section of the gene on the subsequently expressed protein.

Current situation

The researcher searches by gene name, GenBank number in the corporate biology information repository. The then searches GenBank, SwissProt and PubMed for information on this particular gene. In each source, she needs to read the information, find the information describing the protein expression and associated mechanisms of action, and copy the relevant information to a report. This process requires tedious, manual information collection. Specifically, after all this work, it is still unclear whether the information found was up-to-date, or has been succeeded by new information; also, it is unclear whether the status of the experimental evidence is still valid now.

New situation

The COMMIT P23 workflow allows the researcher to obtain all information on the expression on the gene and sections of the gene by using a single query interface. Using her own and pre-existing standard vocabularies, the researcher can quickly formulate her query. The query interface presents the user with the most suitable repositories to query. Using the same interface, the researcher executes reliable, transparent queries over multiple information repositories. Results of these queries are enriched using fine-grained annotations. Based on these annotations, the researcher directly finds the paragraphs associated with her query. She is also given relevant concepts that she can use for further browsing or for reducing the result set. Additional levels of annotation, such as provenance and `epistemic mark-up' provide background to the knowledge claims, linked to experimental evidence, and the argument structure of the article. The researcher is presented with a visualisation of trustworthiness and provenance of annotations, allowing her to list and compare hypotheses. She picks the one that 'holds up' the best, given the experimental evidence and the users' question. An annotation interface allows the researcher to provide additional levels of annotations as she goes along. Researchers can use the same mechanism to enrich their own work.

Requirements

The functionality described above relies on a number of additional advanced processes that take place 'under the hood'. First of all, existing and new data needs to be made continuously accessible through the interface, requiring generic tools for data translation and enrichment. Identity reconciliation is needed to tie statements about the same entities together within and across information repositories. Advanced query mechanisms need to be developed that allow transparent federated querying across (automatically) selected repositories. The structure of semantic annotations needs to be exploited to improve ranking of query results. Provenance of data, and of claims in publications needs to be acquired, represented and connected to the various information sources at suitable levels of granularity (e.g. paragraph-level vs. document-level annotation).

DANS Use Case - Linked Data for e-Humanities

The P23 project will develop a showcase for the DANS collection on the Dutch national census (1795-present). This collection consists, among other resources, of thousands of data tables (currently as spreadsheets), based on the original census publications. This makes it a prototypical example of data-centric heritage that requires translation to the linked data infrastructure developed in P23. The dataset harbours information on demographic, social, economic and cultural characteristics of the Dutch population over the past two centuries, described at the level of neighbourhoods, municipalities, regions, provinces, cities and the rural area. Apart from the thousands of aggregate tables of published census data as spreadsheets (1795-2000), the collection contains a variety of different information types, such as:

- anonymized individual data from censuses 1960-1981,
- images of the digitized census publications and census archives,
- text files and pdf documents of introductory volumes and analytical works on the censuses, and

- an information system on Dutch municipalities, including map coordinates since c. 1815.

The collection is well documented and enriched with extensive metadata. The P23 project envisions both internal links between files within this collection, and external links to linked data resources obtained from P6 and elsewhere. An example is the 2000 U.S. Census, which is available as linked data.[4]

A second use case concerns archaeological data. The E-Depot of Netherlands Archaeology (EDNA) has been the fastest growing area among the DANS data archives and now consists of over 13,000 datasets.[5] Results of archaeological research are often heterogeneous in two different ways: first, the digital documentation of excavations consists of textual, tabular and cartographic documents as well as photographic images. Second, the structure and content of this documentation is geared to the particular research questions concerning the specific archaeological site involved. Within the accepted archaeological methods and techniques there is a lot of room for researchers to document their findings in their own way. In this respect archaeology is representative for much research in the social sciences and humanities.

The consequence of this heterogeneity is that in order to link archaeological data, it has to be harmonised before new research questions can be answered. In the past this harmonization was usually done 'by hand', which was feasible when the datasets were relatively small and few. Due to the so-called Malta-legislation archaeological heritage management has become an integral part of urban planning recently, which has led to an exponential increase in the available digital datasets over the past few years.

A linked data approach by attributing a formal semantics to the data will serve as an important step to make archaeological objects and their context from heterogeneous digital resources interoperable. Formalising typological, spatial and temporal characteristics is necessary for linking different sites on the basis of the current descriptions, which to a considerable extent rely on natural language, and hence are not always as precise as one might wish for.

DANS aims for better integration of scientific publications and (meta) data, and has worked on several 'enhanced publications'.[6] The Research Information (OI) department of the KNAW is in charge of the Dutch Research Database (NOD) with information about scientific research in the Netherlands. Although no full coverage, the NOD provides information about approximately 7,600 professors and senior lecturers, 40,000 researchers and experts, 750 university and non-university research institutes and 120 research schools. The NOD contains 20,000 descriptions of current projects and 18,000 description of completed research. OI also administers the NARCIS portal

[4] See http:/ /www.rdfabout.com/demo/census/ .
[5] Checked on December 15, 2010 at EASY.DANS.KNAW.NL
[6] See, for instance: http://www.dans.knaw.nl/en/node/971; http://www.watveteranenvertellen.nl/; https://www.surfgroepen.nl/sites/JALCproject/default.aspx

offering the combined information from the NOD, more than 600,000 digital publications from academic repositories and some 16,000 data files from the DANS archive.[7] Further integration with other forms of scientific information is in line with the needs of the research community and public funders. No matter what the starting point may be (project, person, data set, publication), it should not be hard to find any corresponding information on each object.

Together with DANS and Elsevier, the P23 project will deliver a specification and showcase for integrating (meta) data in scientific publications using RDFa.

[7] http://www.narcis.nl/

WORKPACKAGES

Project number 23	
WP title & acronym	WP1: Information content and utility
WP leader	Pieter Adriaans, UvA

Objective:

Develop theory and method to measure utility of a given query Q in a given context C over given data D.

Measures of information content and complexity can be used to optimize 'matching' and query utility in a given set of conditions, including context and data.

In WP1 complexity issues concerning queries on large datasets will be studied building on the framework described in Adriaans (2009): Given a certain system S with a certain complexity in the world (i.e. the human brain, climate, DNA, and art style or simply a railroad time table) and a canonical measurement function, i.e. and information channel with certain characteristics that creates a data set D with information 'about' S, under what conditions may we assume that a query Q of a certain form on D indeed returns adequate information about S? In such we will analyze, amongst other things: (1) The conditions under which you can extract 'true' isolated facts from a data set but no general insights, (2) the question whether complex systems that are undersampled create powerlaw distributions (see last point i.e. powerlaws have no means), and (3) the interplay between model information and complexity in the analysis of various systems (i.e. facticity: noise is complex but has a simple model, fractal structures look complex but are simple, lots of structures in nature are both complex and have complex models, specifically products of evolutionary processes)

Project number 23	
WP title & acronym	WP2: Support for querying complex data sets
WP leader	Pieter Adriaans, UvA

Objective:

Develop platform-independent tools that enable retrieval of distributed complex data sets.

Practical tools are needed to provide the means to analyse data utility, and enable uniform access to distributed data resources and their models.

The programmer will help to develop the tools in support of the AIO in WP1 but also help to complete interface and (grid) implementation issues. Important interface issues include: incorporation of configurable vocabularies into the UI, query building tools that help the user to find appropriate terms and completions for a valid query, query federation, transparent mapping of vocabularies and ontological query translation, scoping and access control of vocabularies across organizational boundaries, transparent job farming and parameter sweeping, etc.

Work will result in a mature infrastructure for storage, retrieval, and collaborative annotation of distributed (grid) resources using selected vocabularies. Source code developed will be made public.

Project number 23	
WP title & acronym	WP3: Distance measures and ranking of query results
WP leader	Frank van Harmelen, VU

Objective: Design, development and evaluation of techniques for distance measures and ranking for querying and reasoning on heterogeneous linked data.
Scientific datasets are inherently imprecise, and do not fit neatly into the black and white world of existing Semantic Web technology. Current and nascent techniques for linked data *querying, representation* and *reasoning* must be extended to take semantic distance measures into account.
We will develop reasoning and decision support services that cater for query results and data on a graded or continuous scale. These services can be tailored to meet the demands of different domains and users. This requires fault-tolerant, safe and scalable reasoning over uncertain and incomplete knowledge, embracing data, metadata and knowledge.

Project number 23	
WP title & acronym	WP4: Information publication, integration and enhancement
WP leader	Frank van Harmelen, VU

Objective:

Design, development and evaluation of techniques for publishing, integrating, enriching and reasoning over linked science data.

Scientific data is heterogeneous, both syntactically - it is produced using a variety of tools – and semantically – using a variety of different processes and perspectives. Scientific data must therefore be translated to a unifying paradigm, enriched with the appropriate provenance metadata, interconnected and aligned with shared ontologies, reconciled with data from other datasets, made available for reasoning and querying tasks, and be embedded in scientific publications. This requires extending the current techniques for linked data *publishing* and *enrichment*.

We will develop facilitating infrastructure for linked data management in e-Science, including easy-to-use transformation-generators to cope with legacy formats (in association with LATC), vocabulary linking techniques to cope with semantic heterogeneity, and identity-detection components to cope with co-reference problems. The infrastructure will involve work on large scale reasoning developed in **P20** and LarKC. Requirements for these services will be provided by (non-)profit partners and through synergy with **P6**, **P12**, **P20** and **P26**.

Project number 23	
WP title & acronym	WP5: Information provenance
WP leader	Guus Schreiber, VU

Objective:

Design, development and evaluation of methods and tools for scientific data provenance, covering representing, tracking, annotation and visualization of provenance data.

Provenance data are essential for (re-) interpreting scientific data over time. For correct semantic interpretation of scientific data we need to know, for example, when and how a biological experiment was performed, how a census survey was organized, or when and where a particular historical event description was written down. More generally, the work package aims to provide the appropriate metadata for scientific "objects", such as data sets and articles.

In this WP we base the provenance representation on upcoming standards for the provenance of web data such as the Open Provenance Model[8]. We will develop in an evolutionary fashion methods and tools for generating, maintaining, tracking and interpreting provenance data. These results will be deployed both within **P23**, as well as in the sample domains of application partners in **P6**, **P12**, **P20** and **P26**.

[8] See http://openprovenance.org

Project number 23	
WP title & acronym	WP6: Information annotation and interpretation
WP leader	Guus Schreiber, VU

Objective:

Design, development and evaluation of methods and tools for annotation, interpretation and visualization of scientific data.

In this work package we aim to provide the scientific data analyst with an interactive environment that allows him/her to inspect the results provided by the other work packages.

The toolkit should allow researchers to annotate the results of the data processing tools we provide, adding relevant conceptual links. For this purpose the toolkit should provide multiple visualization tools for showing different semantic views of the data. An example would be to show alternative descriptions of the same historical event (identified through similarities in time and place). The interpretations added by researchers could subsequently be used in further processing, thus leading to a data interpretation workflow.

Project number 23	
WP title & acronym	WP7: Expert-based domain-specific ranking of query results
WP leader	Richard Vdovjak, Philips

Objective:

Design, develop and evaluate new methods for ranking query results that (i) will exploit domain-specific heuristics (in particular in the clinical domain, most likely oncology), and (ii) will exploit "humans in the loop"

All our searches for information have a context and relate to some information or data. For example, clinicians want to extract targeted information that fits the context of a specific patient case (e.g. similar cases reported in literature or stored in a reference database, outcomes for that specific disease, best treatment options, etc.). Current solutions are only able to provide support for very simple questions and decisions, and are not able to fully address the increased complexity of clinical decision for example in the context of oncology.

We will investigate the notion of domain-specific expert ranking, and develop data models and heuristics to describe the notion of expertise, and to associate experts with specific domains. We will develop a system that will enable community-based expert feedback for content and for information sources. We incorporate these expert-driven domain-specific ranking heuristics into the query result ranking in **WP3**. We collect and provide such ranking to enable the extraction of targeted information relevant in a specific domain, for example to support clinically relevant scenarios. We will identify together with clinical experts concrete use cases, such as publication search in a specific clinical domain (most likely oncology), or clinical trial selection based on expert feedback/recommendation.

DELIVERABLES

Number of important journal paper
12

Number of important conference contributions
24

Products

1. Information measures for linked data repositories elf-dissimilarity, saw-tooth phenomena, scale-freeness, MDL-estimates, computational depth, VC-dimension, facticity etc.) An ongoing activity will be the theoretical reflection on measures of meaningful information and their interrelations. In October 2011 we organize a special workshop on measuring meaningful information at the Info-metrics institute in Washington. Results will be made available for P23 and can be a basis for implementation.
   - WP 1 YP 2015

2. Methods for query building
   Platform-independent tools that enable querying of complex data sets. Incorporation of configurable vocabularies into the UI, query building tools that help the user to find appropriate terms and completions for a valid query, query federation, transparent mapping of vocabularies and ontological query translation, scoping and access control of vocabularies across organizational boundaries.
   - WP 2 YP 2013

3. Heuriics for rastnking query results.

This deliverable encompasses an implementation of ranking heuristics based on the requirements set out in Q1 and Q2 of the project. Currently, linked data query engines do not take into account the quality of the results they deliver. We will investigate both quantitative and qualitative requirements for results ranking. The former include requirements related to performance (speed). The latter relate to questions such as: what constitutes a good result, and how can we assess the quality of data and its impact on query result quality, as well as trade-offs between quality and performance in different usage scenarios outlined by the use case partners. We will evaluate the quality of this implementation using P23 as well as COMMIT datasets against the requirements, and compare results to other approaches. Output of this work will be a prototype implementation, as well as further refined requirements for results ranking. Results ranking can be implemented in two distinct ways. First as a static ranking algorithm that will rank results as they are returned from a query engine. The second approach integrates the ranking algorithm and query engine and uses results ranking as dynamic stopping rule. This may lead to more efficient querying over messy data. Depending on what line of research is more likely to produce valuable results, this deliverable will either present a further refined static ranking implementation, and/or an integration of ranking algorithms as dynamic stopping rule in a query engine.

- WP 3 YP 2015

4. Products Methods for integrating data with publications

   This deliverable is an implementation of data integration within publications based on the requirements set out in the first two quarters of the project. We will investigate and identify the requirements of use case partners with respect to the scope and type of data that will be translated, required linkages within datasets as well as with external data (e.g. in the LOD cloud) for identity reconciliation and vocabulary mapping. Current tools for data translation treat the process as a one-off affair, while data enhancement and integration tools require technical skills and do not produce reliable results. We will gather best practices and requirements for publishing scientific data, such as paragraph-level identifiers and the integration of data within publications. Development will follow an iterative approach where the basic functionality (i.e. linking data to publications) is further refined by more elaborate and precise data enrichment technology. This combined effort allows us to link scientific data both from within the P23 project as well as from COMMIT and external parties to publications at the paragraph level. The implementation includes a set of enriched publications as well as a demonstrator for illustrating browsing functionality.

   - WP 4 YP 2013

5. Provenance representations

6. This deliverable will bundle the tools generated at M12 as well as the trust algorithm tools from P6 into a single uniform toolkit. The trust algorithms in P6 will leverage the provenance tools and representations to provide a trust assessment. Importantly, the toolkit will provide a well defined API for managing provenance and using it in visualizations and other algorithms. The toolkit will also leverage existing semantic web infrastructure wherever possible. For example, instead of defining a new query language, SPARQL could be used or extended. The toolkit will describe how to use this existing functionality with respect to provenance. The toolkit will be made available for download. Based on feedback from the continuous evaluation we will update the toolkit. This will entail updating the existing tools but may be adding additional tools. We also believe that the representation may need to be specialized for particular domains and this would be done based on input from various use communities. Importantly, ranking and visualization may put additional demands on the provenance toolkit. This phase will reflect those demands.
   - WP 5 YP 2013

7. Toolkit and visualisation methods for data, queries, results and provenance
   This toolkit will provide a number of visualization tools of scientific data, such as trees, graphs and tables. These will be realized through standard widgets, to prevent as much as possible custom development. Annotations tools will first be of the manual kind, but automatic annotation tools will be added as development progreses, including implementation of alignment methods of WP4. The Amalgame toolkit, built on top of ClioPatria, is expected to provide a starting point.
   The interpretation tools will first implement simple reasoning methods, such discovery of particular, relatively short, graph patterns. At a later stage, interpretation methods will have a wider scope and provide a range of visualizations. A report is included on the first round of evaluation: statistical testing with a large data set, but also qualitative user tests with analyses of think-aloud protocols of user sessions.
   Methods from other WPs will be made available within the toolkit. Work-flow models for particular domains will be developed, that support user interaction. Work-flow models provide typical behavior patterns for e-science research in the domains central to P23. We will report extensive evaluations are reported, including one or more explorations of domains outside the original P23 dataset.
   The final deliverable summarizes the work on the completed toolkit, with an overview of the different tests. Impact on standards is discussed.
   - WP 6 YP 2015

8. Expert-based ranking & inclusion of ranking module in CDS system

This deliverable will be an improved implementation of ranking that makes use of expert knowledge for ranking query results. For this implementation we will focus on data used by clinical decision support systems (CDS), primarily in the domain of oncology. The deliverable will provide an evaluation of this expert-based ranking, and compares results to expectations of domain experts. The expert-based ranking module will be part of the golden demo. We will integrate the ranking into a clinical decision support system, evaluate its performance in several usage scenarios, and compare it to automatic ranking developed in WP3.

- WP 7 YP 2014

Software

1. Golden Demo, 1st version:Assertion-level Document Annotations in Pharmacology

   The functionality of this Demo is described in the project document in section 6.1, p. 12 and concerns the simultaneous navigation of a researcher through publications and research data, while tracking provenance of both. This is a project-wide demo (hence not attributable to a single WP, although WP4 will take the organizational lead). Functionality will combine results from al the WPs:

   - Information measures for linked data repositories (WP1)
   - Query builder & job farming (WP2)
   - Heuristic ranking of results (WP3)
   - Integration of data in publications (WP4)
   - Provenance tracking and trust algorithms (WP5)
   - Visualisation (WP6)
   - Expert-based ranking of results (WP7)
   - Applied to P23 datasets

   - WP 4 YP 2013

2. Golden Demo, 2nd version :Assertion-level Document Annotations in Pharmacology

   This deliverable will be a joint effort by all work packages in presenting an integration of deliverables produced over the last two years. The golden demo will bring together all improved components from each work package. These will be applied and evaluated both against the larger set of P23 data as well as COMMIT datasets.

   WP 4 YP 2015

3. Component prototypes

   Software prototypes will be produced by each WP at every annual milestone (see project document section 9, pg. 18, not repeated here to avoid duplication and inconsistencies). Many of the software prototypes are expected to feed into the Golden Demo.

- WP ? YP 2012

4. Component prototypes
   Software prototypes will be produced by each WP at every annual milestone (see project document section 9, pg. 18, not repeated here to avoid duplication and inconsistencies). Many of the software prototypes are expected to feed into the Golden Demo.
   - WP ? YP 2013

5. Component prototypes
   Software prototypes will be produced by each WP at every annual milestone (see project document section 9, pg. 18, not repeated here to avoid duplication and inconsistencies). Many of the software prototypes are expected to feed into the Golden Demo.
   - WP ? YP 2014

6. Component prototypes
   Software prototypes will be produced by each WP at every annual milestone (see project document section 9, pg. 18, not repeated here to avoid duplication and inconsistencies). Many of the software prototypes are expected to feed into the Golden Demo.
   - WP ? YP 2015

Other results

1. e-Humanities datasets
   e-umanities datasets will be produced based on data available at DANS. These datasets are described at section 6.2, p.15 of the project document
   - WP 4 YP 2012

2. COMMIT Linked Dataset
   A COMMIT Linked Dataset will be produced, containing both scientific data and project metadata (publications, presentations, social network, etc). We will also help other projects to publish their data as Linked Data (arrangement with P26 has been made).
   - WP 4 YP 2015