ontotext

Semantic Technologies in Cultural Heritage

June 2013 Brochure

The Rosetta Stone was key to the deciphering of Egyptian hieroglyphs, by providing parallel text in three scripts: Ancient Egyptian, Demotic and Ancient Greek. Today semantic technologies play a similar role, allowing the Digital Humanist to make connections between (and make sense of) the multitude of digitized cultural artifacts available on the net.

The Digital Humanities embraces and harnesses the expanded, global nature of today's research communities as one of the great inter-disciplinary/post-disciplinary opportunities of our time. It dreams of models of knowledge production and reproduction that leverage the increasingly distributed nature of expertise and knowledge and transform this reality into occasions for scholarly innovation, disciplinary cross-fertilization, and the democratization of knowledge. *– Jeffrey Schnapp, The Digital Humanities Manifesto 2.0*

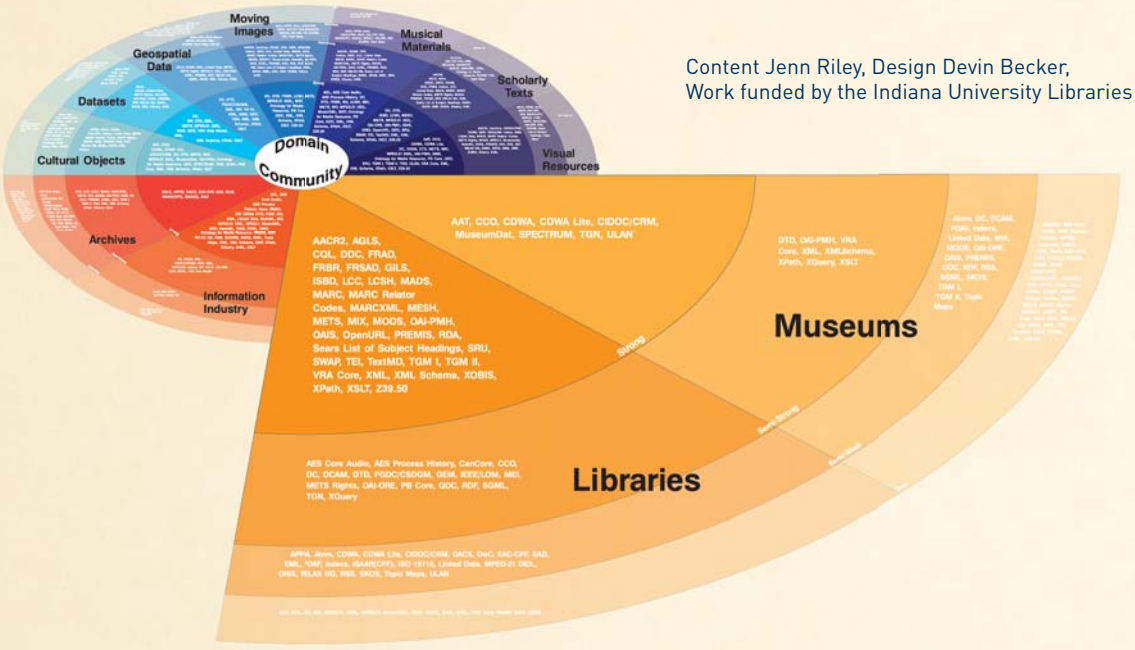
The Digital Humanities Challenge

The internet, global digitization efforts, Europe's Digital agenda, continuing investments in Europeana, the Digital Public Library of America and many other initiatives, have made millions upon millions of digitized cultural artifacts available on the net. We need to make sense of all this information: to aggregate it, find connections, build narratives, analyze data, support the scientific discourse, engage users...

From ancient maps, to bibliographic records, to paintings, to coins and hoards, to paleographic analysis, to proposography factoids... everything is becoming more and more connected. A host of ontologies and metadata standards have come into existence: CIDOC CRM, TEI5, LIDO, SPECTRUM, VRA Core, MPEG7, DC, ESE and EDM, OAI ORE and PMH, the list goes on and on.

A number of established thesauri and gazetteers exist, some of them interconnected: DBPedia; VIAF, FAST, ULAN; GeoNames, Pleiades, TGN; LCSH, AAT, IconClass, Joconde, SVCN, Wordnet, etc etc. How to use them in every-day collection management, cataloging, documentation and research?

Below is **a small part** of the "Seeing Standards" poster: **are you seeing double?**



Content Jenn Riley, Design Devin Becker,
Work funded by the Indiana University Libraries

How to preserve the role of libraries, museums and other Cultural Heritage institutions as **centers of wisdom and culture** into the new millennium? Aren't Google, Wikipedia, Facebook, Twitter and smart-phone apps becoming the new centers of research and culture (or popular culture at least)?

We believe the answers to all these questions lie with **Semantic Technologies**. They enable large-scale Digital Humanities research, collaboration and aggregation; and technological renewal of Cultural Heritage institutions

Ontotext's Cultural Heritage Experience

Ontotext is a well-known world-class developer of core semantic technologies, including Semantic Repositories, Text Analysis, Information Extraction and Retrieval, Semantic Annotation and Search, Media Publishing, Web Mining, Information Integration.

A couple of years ago we started working in cultural heritage, and have accumulated significant knowledge, experience, projects, clients and partners. The Data and Ontology Management group of Ontotext has two mandates:

- technical: metadata standards, data conversion and integration, ontologies, thesauri
- vertical: applications in the cultural heritage domain

We hope to collaborate with you on your next project, or indeed to define it together!

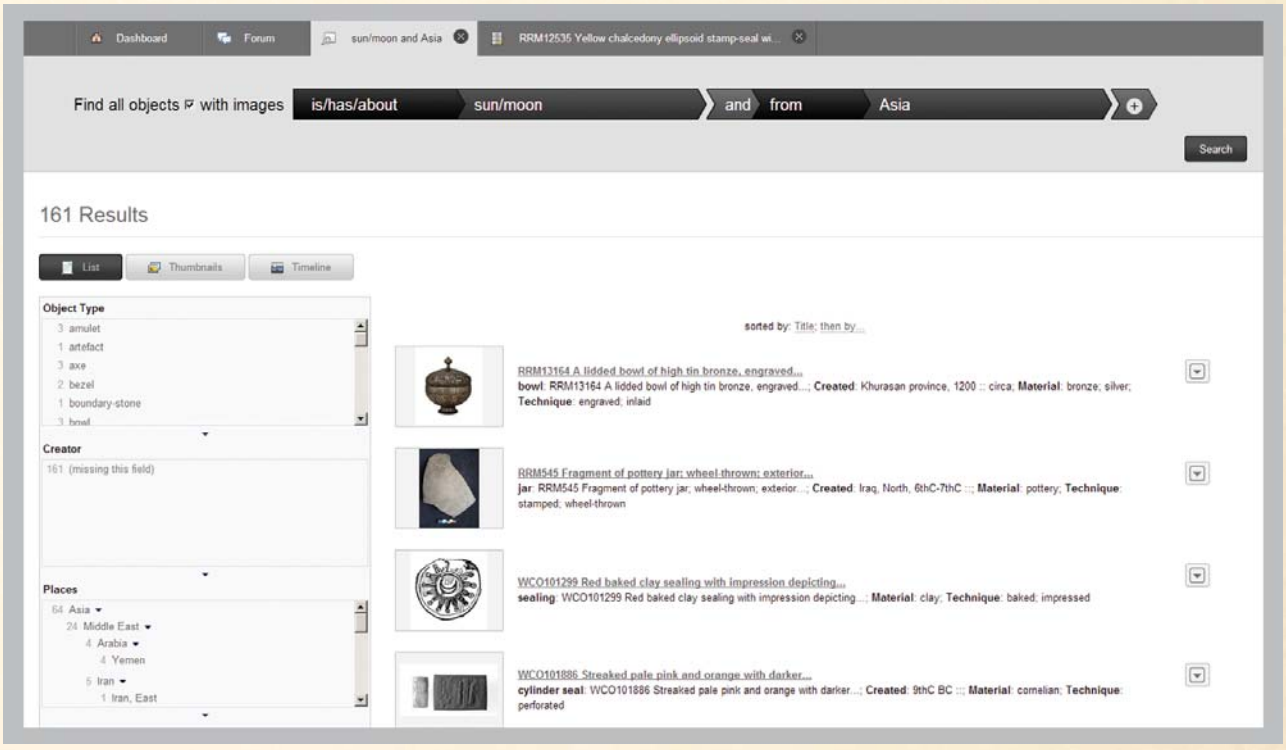
ResearchSpace: Semantic Search

ResearchSpace (RS) is a project of the British Museum (London), funded by the Andrew W. Mellon Foundation (USA). It aims to support collaborative research projects for cultural heritage scholars. RS will implement an open source framework and hosted environment for web-based research, knowledge sharing and web publishing. RS intends to provide:

- Data conversion and aggregation
- Semantic RDF data sources, based on the CIDOC CRM ontology
- Semantic search based on Fundamental Relations
- Data analysis and management tools
- Collaboration tools, such as forums, tags, data baskets, sharing, dashboards
- Various research tools and workflows, e.g. Image Annotation, Image Compare, Timeline and Geographical Mapping...
- Web Publication

Semantic technology is at the core of this project because it provides an effective mechanism for research and collaboration across data from different organizations and projects. RS uses Ontotext's OWLIM semantic repository featuring powerful reasoning (equivalent to OWL2 RL), fast performance, efficient multi-user access, full SPARQL 1.1 support, and incremental assert and retract.

RS Stage 3 (Working Prototype) was developed between Nov 2011 and Apr 2013 (software development by Ontotext). Development of Stage 4 is expected to start in 2013, with more museums and galleries coming on board.



The Collections Trust will be working with the British Museum to explore the implications of this new Create Once, Publish Everywhere (COPE) approach, and to share it as widely as possible with the museum, gallery and built heritage communities.
ResearchSpace is an interesting case in point - it is, at heart, a linked open data documentation system on steroids. But its look and feel wouldn't be out of place in a high-end enterprise application... An environment which is neither front-of-house, nor back-office, but both at the same time. It does a hardcore, complex museum job, but it does it in an environment which would (I think) feel as comfortable for a casual user as it would for an academic researcher or expert curator.

– Nick Poole, CEO, Collections Trust

I have started demonstrating the tools in the ResearchSpace working prototype and received very positive feedback. The contextual nature of the search system, using the rich descriptions and harmonisation properties provided by the CIDOC CRM, has been particularly well received and provides avenues for exploring data in entirely new ways.

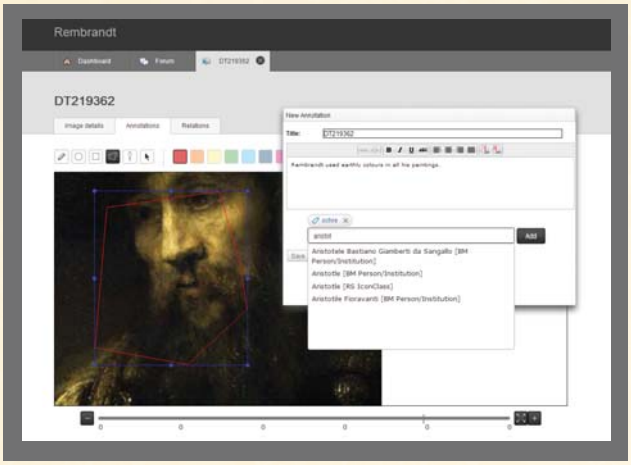
– Dominic Oldman, IS Development Manager, the British Museum and ResearchSpace Principal Investigator

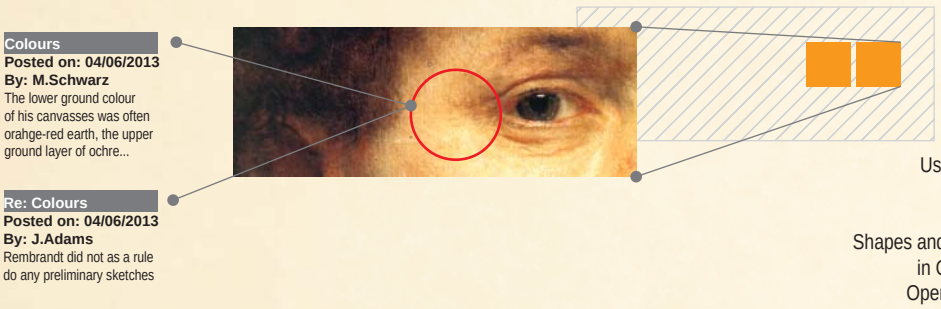
See RS semantic search in action:
http://www.youtube.com/watch?v=HCnwgq6ebAs

ResearchSpace: Image Annotation

The Image Annotation tool provides core functionality for collaborative research on paintings and high-resolution photos of objects. Features:

- Draw arbitrary shapes over an image (most open-source annotation tools allow only rectangular shapes). We use the open-source library SVG-Edit. Scalable Vector Graphics (SVG) supports shapes, colors, line styles, markers and more.
- Deep Zoom support for high-resolution (multi-gigapixel) images. We use the open-source IIP Image Server. Annotations can be created at any zoom level, and are scaled accordingly at different levels
- Attach any semantic object, comment, reply or threaded discussion to shapes
- Image overlay and blending (limited version, to be extended)
- Uses the OpenAnnotation data model (another Mellon-funded project)





Deep zoom server serves tiles

to a JavaScript image viewer

Users draw any annotation shape using SVG-Edit

Shapes and annotations are saved in OWLIM triple store using Open Annotation data model

CIDOC CRM

The CIDOC CRM is a compact top-level (conceptual) ontology that is appropriate for cultural heritage, historic discourse, archaeology. It supports generic description of cultural artifacts, people, places, sites, related events (e.g. creation, acquisition, finding, curation, conservation), cultural periods. Ontotext helped the British Museum to develop its mapping to CIDOC CRM, and Best Practice guidelines that other museums can use. We promote CRM extensions and corrections that facilitate real interoperability and federation between collections of different institutions.

Ontotext organizes the workshop **Practical Experiences with CIDOC CRM and its Extensions** (CRMEX 2013) at TPDL 2013 (26 Sep 2013, Malta)

CRM Reasoning, Search, Performance

Ontotext provided the first working implementation of CRM Fundamental Relations (FRs) search over large data. We used OWLIM Rules, which provide equivalent reasoning power to OWL2 RL, and efficient incremental updates. We implemented 20 FRs using 104 rules and about 40 sub-FRs.

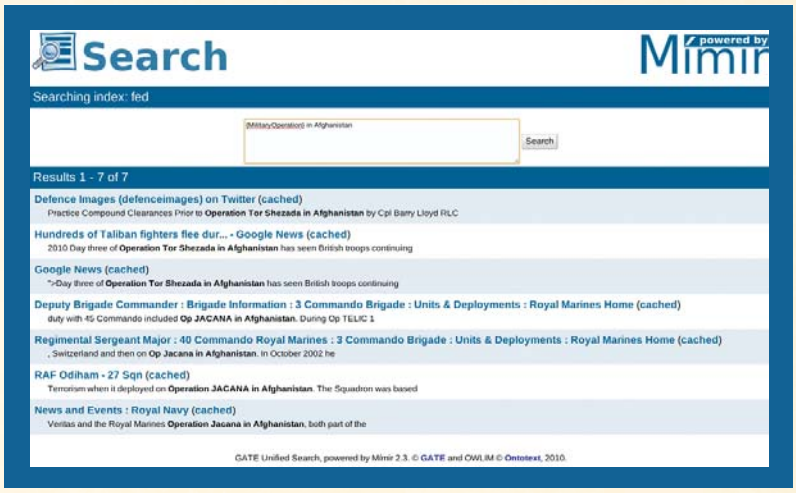
The BM data in RS comprises **2M museum objects, 53M RDF nodes, 194M explicit statements, and 1.5B total statements.** This means that each explicit statement generates about 7 statements, inferred through forward chaining and stored using materialization. This high ratio of inferred statements is due to the deep class hierarchy of CRM (about half of all statements are rdf:type), transitively closed and inverse properties; the search FRs generate about 6% of all statements.

Despite this large amount of data, OWLIM provides good search response times. This is an exciting demonstration of large-scale reasoning with real-world data: no other repository has demonstrated such expressive reasoning with more than 5-10M synthetic statements.

The National Archives

Another example of semantic processing and semantic search at extreme scales is the TNA Semantic Knowledge Base project. TNA has been archiving all central UK government web sites since 2007 (key websites since 1997), amassing a wealth of information. But the users had no good way to search through this huge archive. Ontotext:

- Received from TNA some 700M documents, comprising 1.3B files in 260K compressed archives, and 42TB uncompressed
- Performed de-duplication, reducing the set to 160M unique documents
- Loaded the complete UK Government Ontology to OWLIM: including structure, departments, offices, positions, office holders, and historic evolution. A total of 2B explicit and 5B total facts (triples).
- Performed text analysis and semantic annotation for automatic text classification by topic and document type, named entity extraction
- Created a multi-paradigm semantic search system. Eg below is a conceptual search ("Military operation") involving context ("in Afghanistan"):

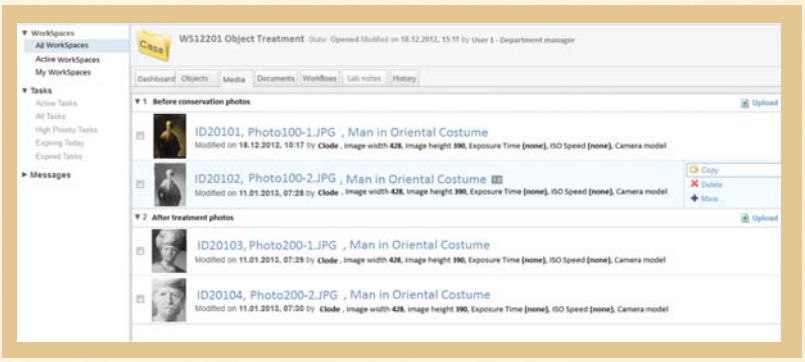


The Semantic Knowledge Base implementation delivered by Ontotext solves two problems: the collection contains a large amount of near-duplicate material which makes conventional search tools not very effective; another major issue is how users continue to find topical information when government structures change over time.

– Amanda Spencer
Head of Web Continuity
The National Archives
UK

ConservationSpace

ConservationSpace is another project funded by the Andrew W. Mellon Foundation (USA). It is managed by the National Gallery of Art (USA) and 7 other institutional partners from the USA, UK and Denmark.
It will develop an open-source application to address a core need of the conservation community: a shared solution to the problem of documentation management. The partners have invested four years since 2009 to define an extensive set of use cases, work processes, requirements, and document descriptions.



Sirma ITT and Ontotext won the international tender for the Build phase of the project, making this the second Mellon project of Ontotext. Work started in Mar 2013 with a Discovery Phase. The system will be based on Alfresco document and case management, and likely on semantic technologies and OWLIM. Later phases will include integration to Collection Management Systems and to ResearchSpace.

Europeana, EDM endpoint, Bulgariana

Europeana is the pan-European digital library and museum. In Oct 2012 Europeana opened for free download its complete RDF dataset in EDM (Europeana Data Model) format, comprising over 20M cultural heritage objects from across Europe. Ontotext was invited to load the data and host the Europeana SPARQL endpoint.

The data is loaded in the OWLIM semantic repository with OWL-Horst inference, and using Ontotext's Forest framework as a front end. It comprises 993M explicit and 4B retrievable statements and is accessible at http://europeana.ontotext.com. Although still experimental, it received a warm welcome from the semantic web community.

This is the first large public endpoint that I've seen with SPARQL 1.1 support. I didn't need any SPARQL 1.1 features for the query above, but did for others on my way there - for example, to find out that there were 6,219 audio files... The same query can be easily modified to return other types of content available from Europeana, e.g. video. As a SPARQL geek's alternative to YouTube, the 166,872 resources with an edm:type value of "VIDEO" are a tempting way to kill some time

– Bob DuCharme,
semantic solution architect,
TopQuadrant

The Europeana Creative project will demonstrate that Europeana can facilitate the creative re-use of cultural heritage metadata and content. The project will establish an Open Laboratory Network, create a legal and business framework for content re-use and implement all needed technical infrastructure. The project will create five pilot applications in the thematic areas of History Education, Natural History Education, Tourism, Social Networks, and Design, then conduct open innovation challenges to identify, incubate and spin-off viable projects into the commercial sector.

Ontotext is involved in core backend tasks: triple store as a central data integration repository, and Content Retrieval Services as part of the Content Re-use Framework. Ontotext will also work on geo-referencing of metadata, Geographic mapping and transformation algorithms, and linking to external web resources. As part of the project, Ontotext will also develop further the EDM semantic repository, improve performance, create custom views to display the data better, and maintain it.

Europeana Creative involves 26 organizations ranging from museums and libraries, to innovation and creative hubs, to proven technical partners. This is our first CIP PSP project. It is a collaboration with a host of excellent partners, and we hope it will be followed by other useful collaborations.

Ontotext and Sirma Media established the Bulgariana aggregator, which contributes Bulgarian content to Europeana. The initial work was performed in a project of the Bulgaria-Korea IT Cooperation Center, but this is an open-ended initiative whose mission is to invigorate digitization and semantic publication of Bulgarian cutural heritage.

We initially published some key Bulgarian artefacts, and are currently in discussions with Bulgarian museums to do more.

Clients

The UK National Archives (London): Semantic Knowledge Base

The British Museum (London): ResearchSpace project

National Gallery of Art (Washington USA): ConservationSpace project

FP7 Europeana Creative: re-use of cultural heritage metadata and content by the creative industries

Europeana: EDM semantic data SPARQL endpoint

Bulgaria-Korea IT Cooperation Center:
semantic publishing of key cultural heritage collections
Bulgariana: aggregator to contribute Bulgarian content to Europeana

Yale Center for British Art (USA): Linked Open Data publishing of museum collection

Dutch Public Library (Netherlands): cultural heritage aggregation

Gothenburg City Museum (Sweden): semantic presentation of collection, natural language question answering

National Institute of Informatics (Japan): LODAC

Polish Digital National Museum: aggregates 700 thousand objects

FP7 MOLTO: Multilingual Online Translation, with application to museum collections

FP7 CHARISMA: Synergy for a Multidisciplinary Approach to Conservation/Restoration

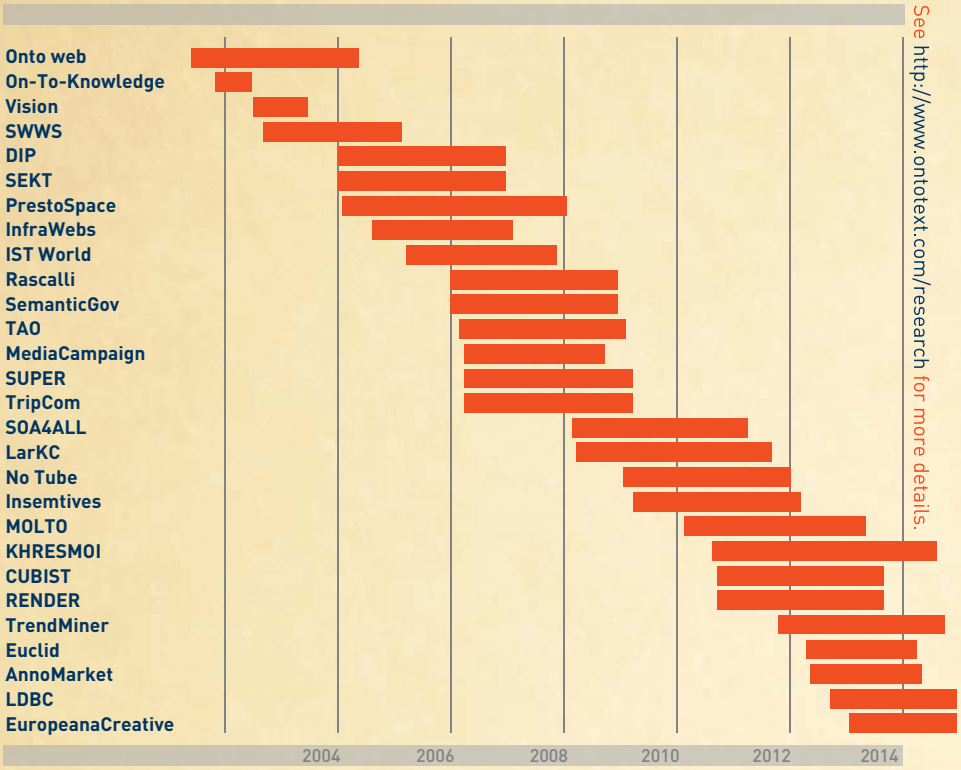
FP7 3D-COFORM: 3D documentation and collection formation of tangible cultural heritage

FP7 V-MUST:
Virtual Museum Transnational Network, a Network of Excellence

FP7 Idea Garden:
learning environment to assist designers during all phases of the creative process

See **http://www.ontotext.com/clients#culture** for more details

Ontotext has participated in 30 EU projects since 2002.



See http://www.ontotext.com/research for more details.

MEMBERSHIPS

Ontotext is part of Sirma Group

STI · INTERNATIONAL

Current research projects relevant to Cultural Heritage:

MOLTO - Multilingual Online Translation - developing tools for translating texts between multiple languages in real time with high quality. Ontotext leads a Museum use case for the Gothenburg City Museum

RENDER - Reflecting Knowledge Diversity - developing methods, techniques, software and data sets that will leverage diversity as a crucial source of innovation and creativity. Techniques developed together with Google for relating news articles to Linked Open Data, and for clustering entities, can be used profitably on CH data.

EUCLID - Educational Curriculum for the usage of Linked Data - professional training curriculum for data practitioners aiming to use Linked Data in their daily work. Strongly relevant to cultural heritage metadata specialists and other experts focusing on Linked Open Data.

AnnoMarket - Cloud-Based Text Annotation Marketplace - aims to revolutionize the text annotation market, by delivering an affordable, open marketplace for pay-as-you-go, cloud-based extraction resources and services, in multiple languages. Multilingual semantic entity extraction from cultural heritage text (e.g. museum object descriptions) is an important and largely unsolved problem. Ontotext's strong experience in this domain, as well as this particular project, provide important avenues for addressing the problem.

LDBC - Linked Data Benchmark Council - aims to establish a global, vendor-neutral, non-profit organization for publishing and auditing benchmark results for graph and RDF databases. Cultural heritage institutions that want to use semantic repositories require benchmark information, and can provide important cultural heritage data sets, use cases and feedback.

Europeana Creative - re-use of cultural heritage metadata and content by the creative industries. This project provides a crucial contribution to improving the usefulness and kick-starting the professional use of Europeana data. Ontotext plays a core role in the heart of the developed system, namely the Content Re-use Framework. This is our first CIP PSP project. It is a collaboration with a host of excellent partners, and we hope it will be followed by other useful collaborations.

Ontotext AD
47A Tzarigradsko Chaussee
Floor 4
1504 Sofia
Bulgaria

Tel: (+359 2) 974 61 60
Fax: (+359 2) 975 32 26

info@ontotext.com
www.ontotext.com