VCMS Project & Proposal Design

Vladimir Alexiev, PhD, PMP*
Data and Ontology Group, Ontotext

(*what is this TLA? May have bearing later on…)

COST Action IS1005, Medioevo Europeo
VCMS Meeting, Budapest, Hungary, 17-Oct-13

Presentation Outline

- About Ontotext
  - Clients
  - Projects in Cultural Heritage
  - A large-scale project
  - European projects
- Sample Projects in eInfrastructures and Digital Humanities
- VCMS Proposal, Project, System Considerations
- EU Horizon 2020 Funding Instruments
  - FET Open
  - ICT 2014-2015
  - eInfrastructures
  - Humanities??

About Ontotext

- Innovative BG SME, leader in Semantic Technology software
  - Semantic database (repository): OWLIM
  - Text analytics, semantic annotation, entity extraction, search: KIM
  - Web mining: job offers, cars, recipes, etc.
  - Data integration, conversion, metadata and ontology management, Linked Data
- Verticals and markets
  - Publishing (dynamic semantic publishing), both Media and Publishers
  - Life Sciences and pharmaceuticals
  - Cultural Heritage and Digital Humanities
- Company information
  - Part of Sirma Group, largest private Bulgarian software holding
  - Established in 2000 as a research laboratory, working on NLP and semantics
  - Received venture funding and spun off as separate company in 2008
  - 80 employees and contractors
  - Offices in Bulgaria (Sofia, Varna). Representation in London and New York

ontotext

Ontotext Clients

BBC

PRESS ASSOCIATION

NDP Nieuwsmedia
De brancheorganisatie voor nieuwsbedrijven

The National Archives

THE BRITISH MUSEUM

WILEY
Publishers Since 1807

OXFORD
UNIVERSITY PRESS

Natural Resources Canada

YALE CENTER FOR BRITISH ART

LMI
COMPLEX PROBLEMS. PRACTICAL SOLUTIONS.

SIEMENS

kt

South London and Maudsley NHS
NHS Foundation Trust

RJ LEE GROUP

GÖTEBORGS STADSMUSEUM

AstraZeneca

legislation.gov.uk

GOVERNANCE
BASEL INSTITUTE ON GOVERNANCE

FAO
FIAT PANIS

Raytheon

CYFROWE MUZEUM NARODOWE W WARSZAWIE

HOUSES OF PARLIAMENT
INFORMATION & COMMUNICATIONS TECHNOLOGY

EUROMONEY

europeana
think culture

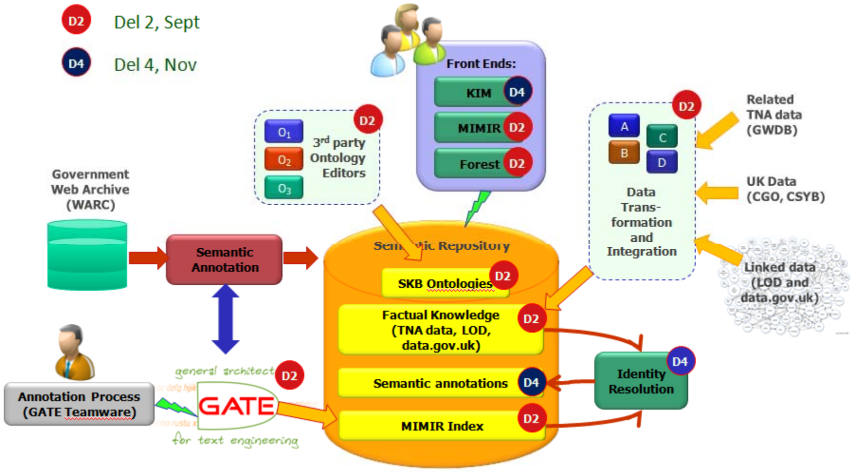
Projects in Cultural Heritage

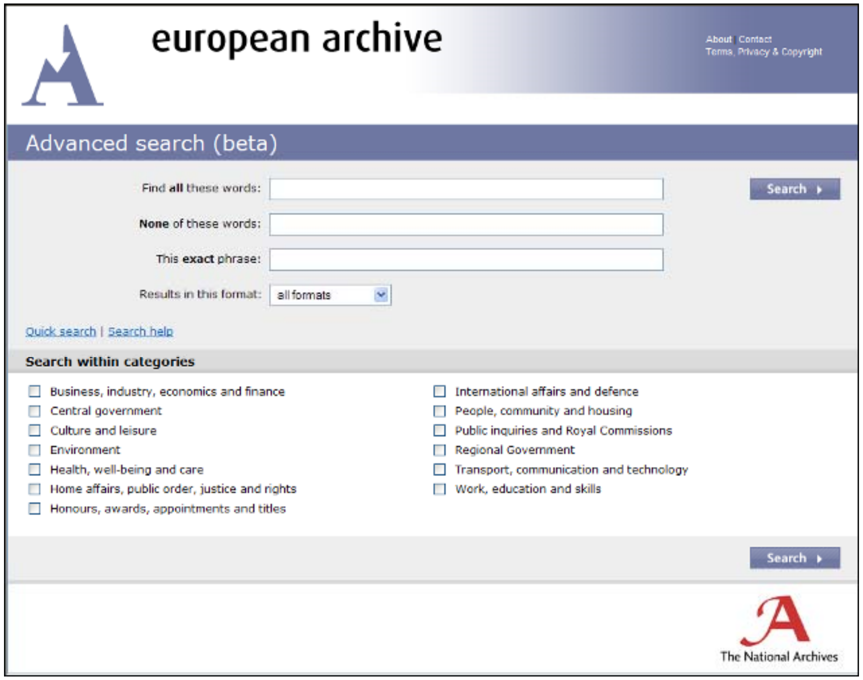
- **The National Archives** (UK): Semantic Knowledge Base
- **The British Museum** (UK): ResearchSpace project, funding by Mellon Foundation
- **Yale Center for British Art** (USA): Linked Open Data publishing of museum collection
- **National Gallery of Art** (US): ConservationSpace project, funding by Mellon Foundation
- **Bulgariana**: aggregator to contribute key Bulgarian content to Europeana
- **Europeana EDM SPARQL endpoint**: http://europeana.ontotext.com
- **Europeana Creative**: re-use of cultural heritage metadata and content by the creative industries
- **Ambrosia** (Europeana Food and Drink): explore and celebrate European cultural identity through its culinary and social history.
- **Dutch Public Library** (Netherlands): cultural heritage aggregation
- **Projects using Ontotext technology**: 3D COFORM, V-MUST, IdeaGarden, CHARISMA, LODAC, Polish Digital National Museum…
- Active in **CIDOC CRM** (organized CRMEX workshop on practical experience with CRM)

ontotext

UK National Archives: Semantic KB

- Example of large-scale semantic processing
- Semantic index for the entire UK Government Web Archive
- **700M** documents: 42TB, 1.3B files
- **160M** unique documents after de-duplication
- Background knowledge (UK Government Ontology): **5B facts**
- Automatic text analysis: **extracted 3B facts** of metadata
- Faceted semantic search in KIM
- **33K hours** of cloud processing; up to 500 servers
- www.ontotext.com/case/nationalArchives-skb





european archive

About Contact
Terms, Privacy & Copyright

Advanced search (beta)

Find **all** these words: ______ Search ▸

None of these words: ______

This **exact** phrase: ______

Results in this format: all formats

Quick search | Search help

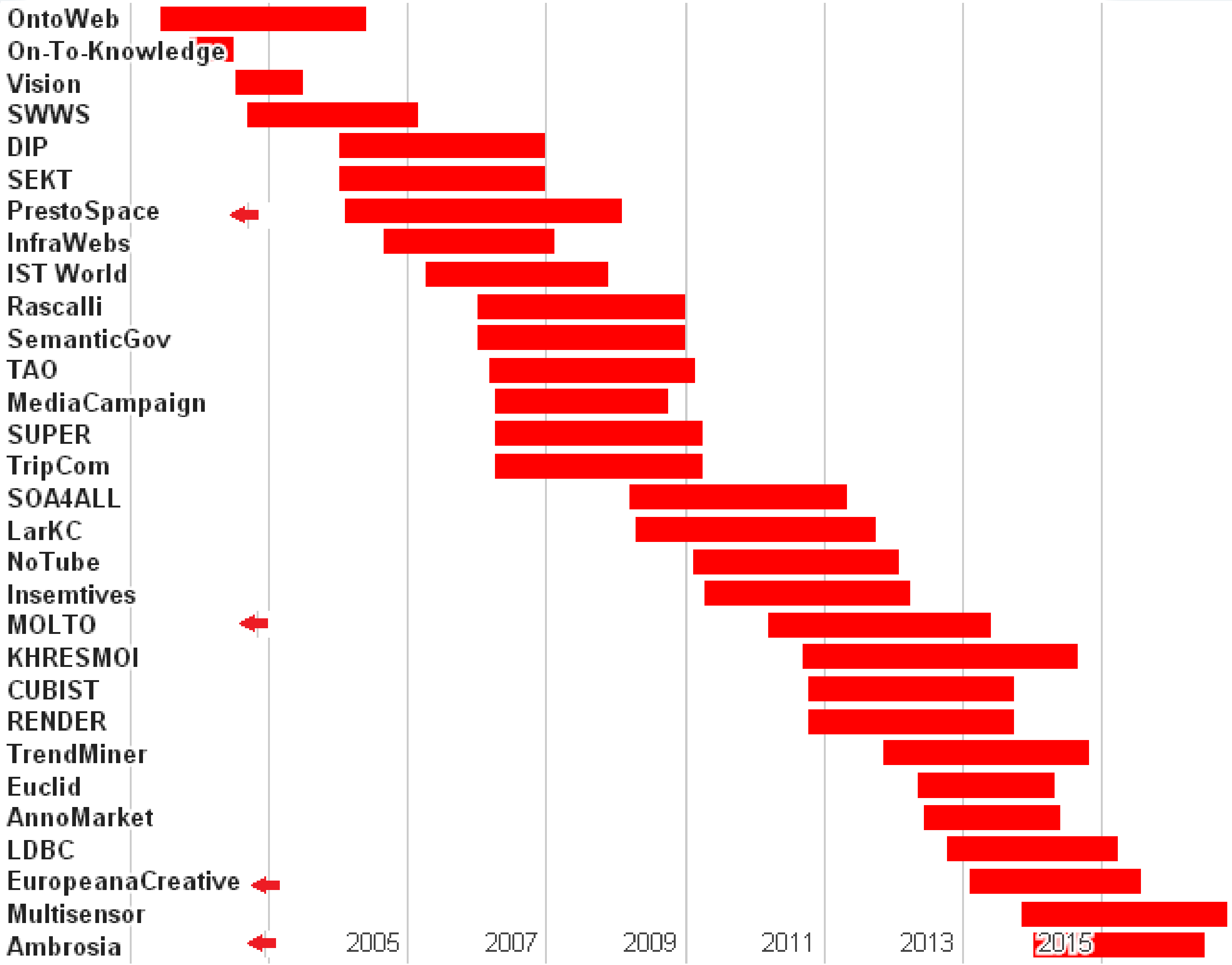
Search within categories

□ Business, industry, economics and finance	□ International affairs and defence
□ Central government	□ People, community and housing
□ Culture and leisure	□ Public inquiries and Royal Commissions
□ Environment	□ Regional Government
□ Health, well-being and care	□ Transport, communication and technology
□ Home affairs, public order, justice and rights	□ Work, education and skills
□ Honours, awards, appointments and titles	

Search ▸

The National Archives

EC Research Projects (FP5-FP7)

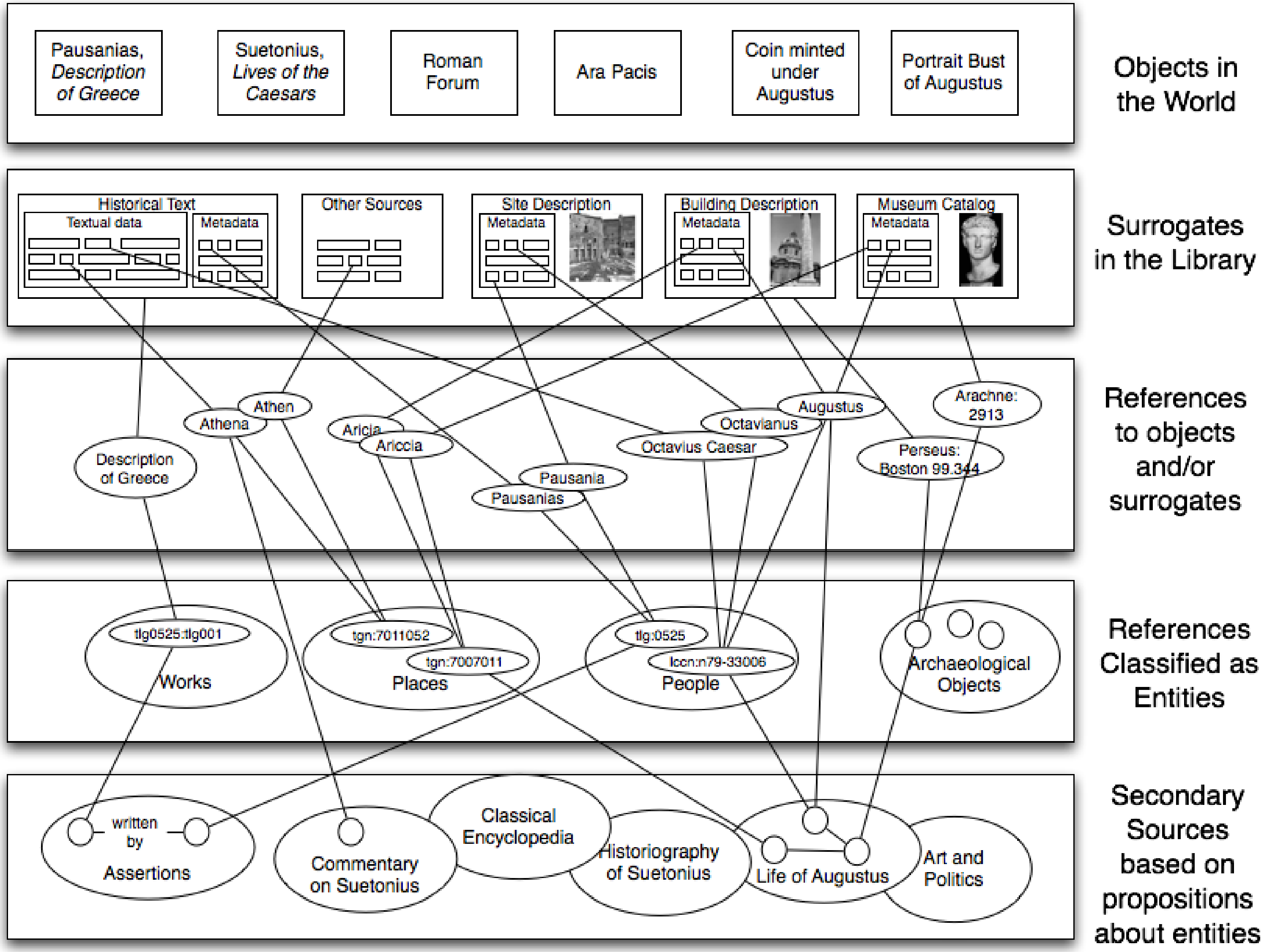


- Bulgaria's biggest participant. 20.8% of projects (15 of 72), 36.6% of funding
- Arrows: CH projects. About ~10 more projects are also relevant to CH

Research Infrastructures, Digital Humanities…

SAMPLE PROJECTS

Model of Digital Scholarship



Robert Kummer, "Named Entity Identification / Disambiguation", Uni Koeln, Sep 2007

ontotext

Research Infrastructures in the Humanities

- RI are about data centers, peta-bytes, mega-FLOPS, millions of cores…
- Are RI relevant to Humanities?
  RI in the Digital Humanities:
  ESF Science Policy Briefing 2011
  4 Definitions, Taxonomies and Typologies of RIs
  8 Bridging Physical RIs in the Humanities with Digital RIs
  15 Researchers' Input and Engagement in Producing RIs
  19 Digital Research in the Humanities: who is Responsible for RIs?
  21 Preservation and Sustainability
  24 Evaluation of Digital Research and its Outputs
  26 Communities of Practice
  32 Cultural and Linguistic Variety – Transnational RIs
  35 Education and Training
  39 Conclusions: Priorities for Policy and Research
  42 References
- CLARIN (ERIC), DARIAH (getting there)

European Science Foundation

42 Science Policy Briefing September 2011

Research Infrastructures in the Digital Humanities

Executive Summary

42 Science Policy Briefing • September 2011

Research Infrastructures in the Digital Humanities

Contents

ontotext

Sample RI project: ARIADNE

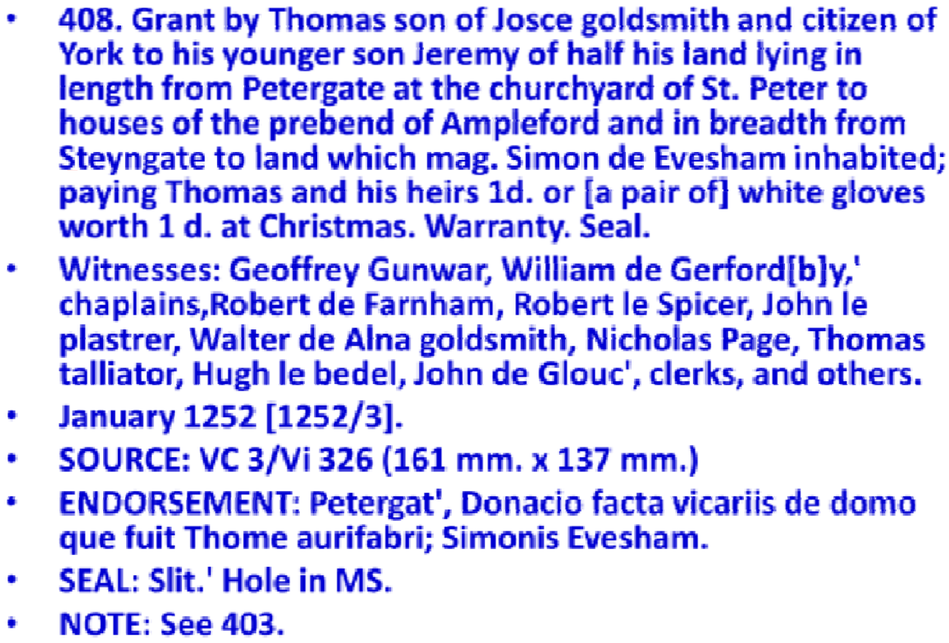
- Advanced Research Infrastructure for Archaeological Dataset Networking in Europe
- Associated to (but not funded by!) DARIAH
- Funding: FP7 eInfrastructures IA
- EU financial contribution: 6.5 MEUR
- Period: 48 months (Feb 2013 - Jan 2017)
- 24 partners from 13 countries. Include most existing national services, e.g.: ADS UK, SNDS, DANS NL, DAI DE, Fasti Online

Sample Project: ChartEx

- Funded by the Digging Into Data scheme
  - 4 countries, 10 funding agencies
- New ways of exploring full text content of digital historical records: *medieval charters.* One of the richest sources for studying the lives of people in the past (prosopography)
- Enable users to *dig into the data* of these records, to recover their rich descriptions of *places and people*
- Go beyond current digital catalogues which restrict searches to a few key facts about each document (the 'metadata')
- Uses machine learning NLP techniques: learns from a manually curated Gold Standard corpus
  - Same approach was we use with commercial Concept Extraction services for Media and Publishers: BBC, UK Press Association, NDP Dutch Press Association, etc
- Uses BRAT, a text annotation tool that allows to express entities, concepts, relations, sentence structure, metaphor, etc
  - We also intend to use BRAT in our projects

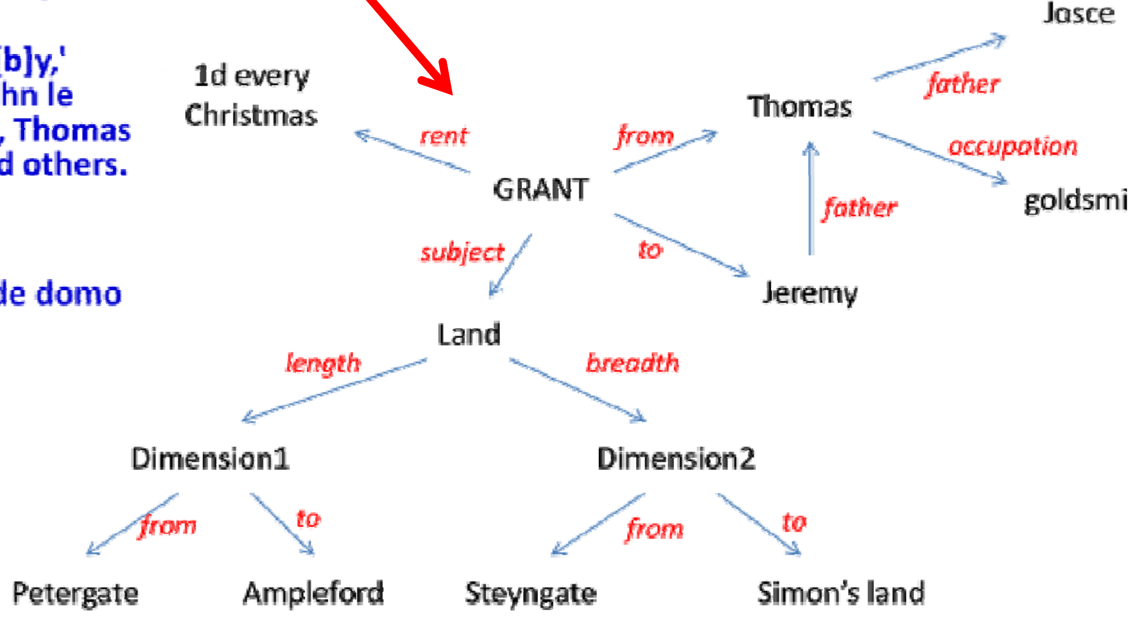
ChartEx Purpose

- From charters to data (networks of related entities)



- 408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and his heirs 1d. or [a pair of] white gloves worth 1 d. at Christmas. Warranty. Seal.
- Witnesses: Geoffrey Gunwar, William de Gerford[b]y,' chaplains,Robert de Farnham, Robert le Spicer, John le plastrer, Walter de Alna goldsmith, Nicholas Page, Thomas talliator, Hugh le bedel, John de Glouc', clerks, and others.
- January 1252 [1252/3].
- SOURCE: VC 3/Vi 326 (161 mm. x 137 mm.)
- ENDORSEMENT: Petergat', Donacio facta vicariis de domo que fuit Thome aurifabri; Simonis Evesham.
- SEAL: Slit.' Hole in MS.
- NOTE: See 403.

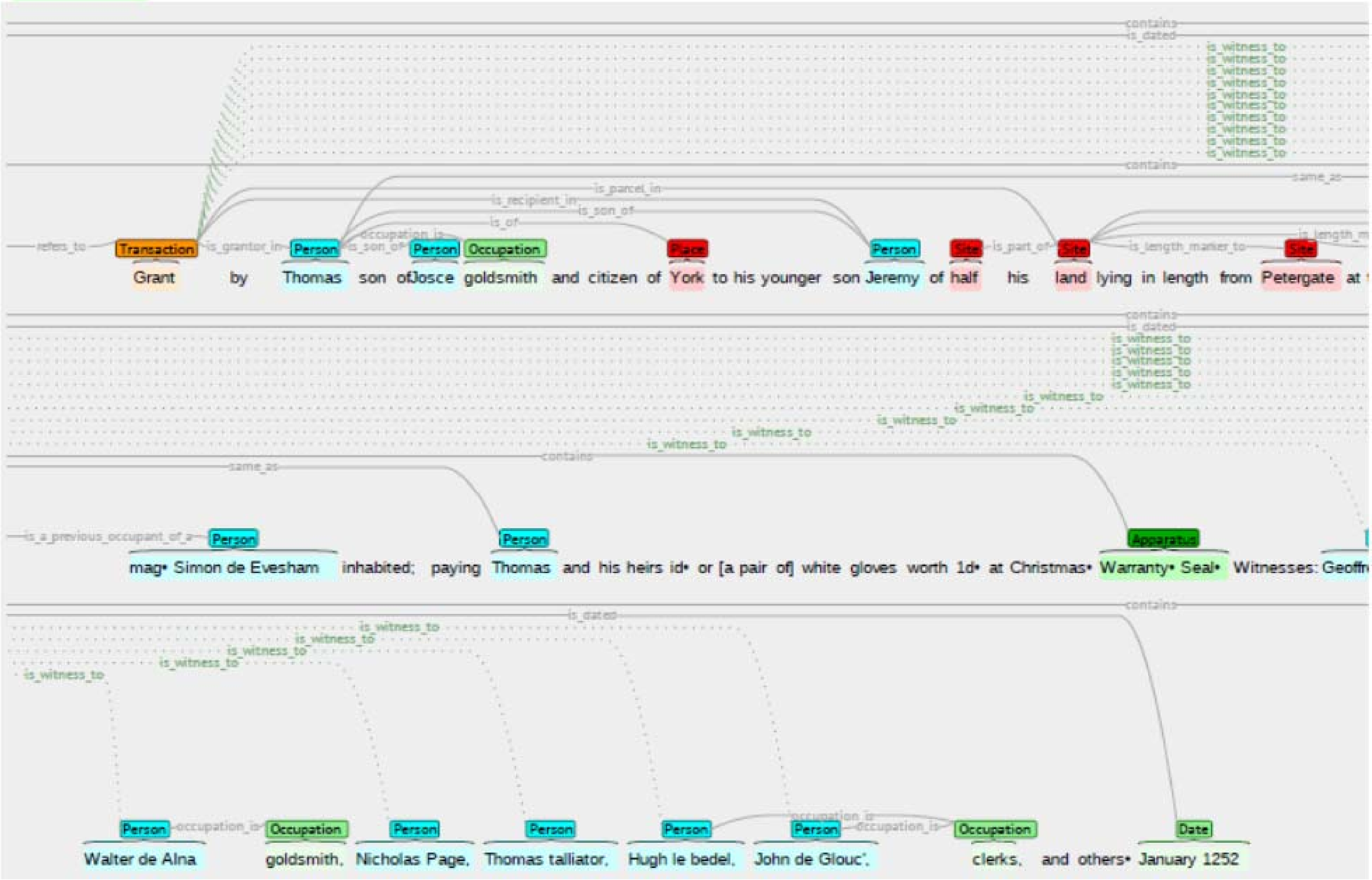
how?



- What for?
  Extract **history** from documents

- Sarah Rees Jones, Christopher Power;
  ChartEx: Discovering spatial descriptions and relationships in medieval charters.
  University of York, Nov 2012

Use of BRAT in ChartEx



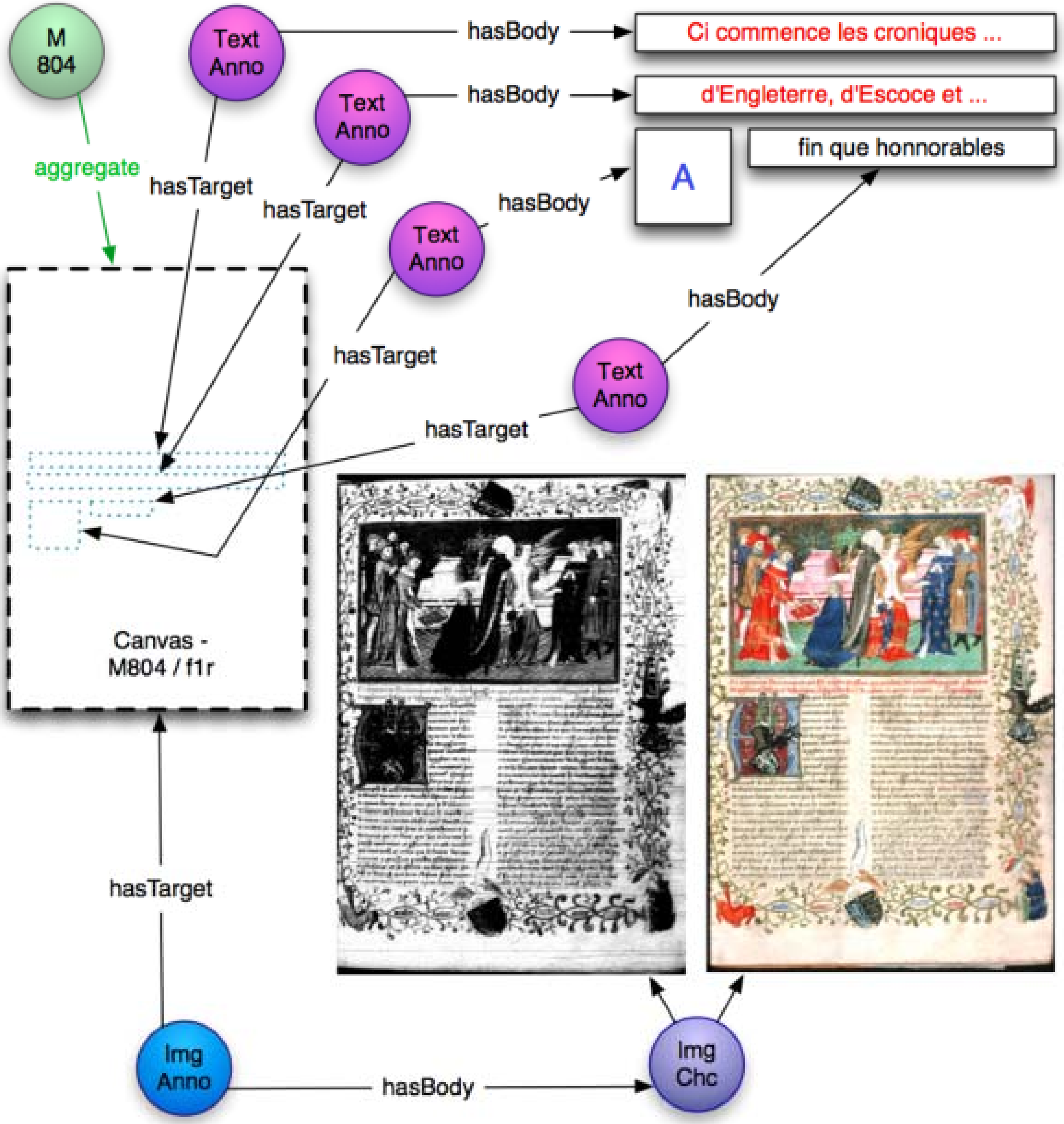
Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at

mag• Simon de Evesham inhabited; paying Thomas and his heirs id• or [a pair of] white gloves worth 1d• at Christmas• Warranty• Seal• Witnesses: Geoffr

Walter de Alna goldsmith, Nicholas Page, Thomas talliator, Hugh le bedel, John de Glouc', clerks, and others• January 1252

Sample Project: SharedCanvas

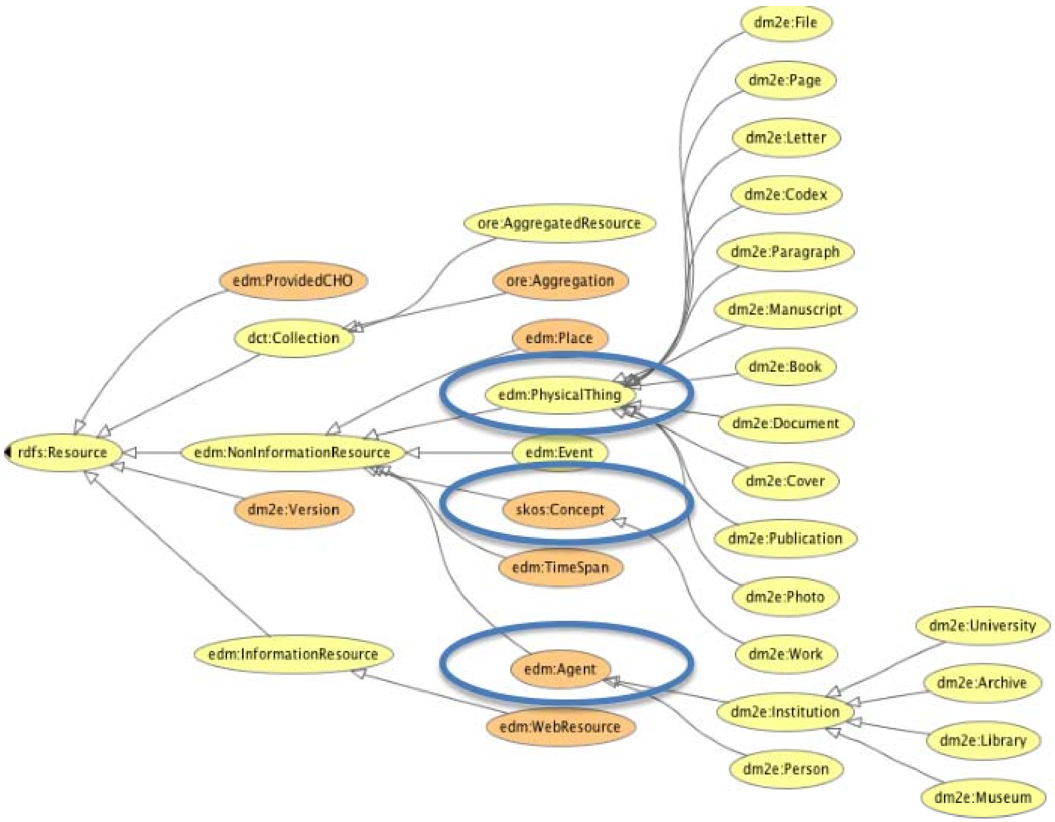
- SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination
- Robert Sanderson, Herber Van de Sompel (LANL), Benjamin Albritton (Stanford U), Rafael Schwemmer (U Fribourg)
- OpenAnnotation & AnnotationOntology: unified ontology for annotation, bookmarking, placement of texts, images, videos…



ontotext

Sample Ontology: DM2E EDM+

- "Digital Manuscripts to Europeana" project
- Ontology for manuscripts (extension of EDM), conversion of sources
- Use OWLIM as semantic repository



🟦 dm2e:copyist
🟦 dm2e:copyrightHolder
🟦 dm2e:dateOfRemark
🟦 dm2e:editor
🟦 dm2e:hasVersion
🟦 dm2e:honoree
🟦 dm2e:influencedBy
🟦 dm2e:isPartOf
🟦 dm2e:locatedIn
🟦 dm2e:mentioned
🟦 dm2e:misattributed
🟦 dm2e:owner
🟦 dm2e:portrayed
🟦 dm2e:printer
🟦 dm2e:recipient
🟦 dm2e:refersTo
🟦 dm2e:sponsor
🟦 dm2e:studentOf
🟦 dm2e:subjectOf
🟦 dm2e:translator
🟦 dm2e:wasStudiedBy
🟦 dm2e:wasTaughtBy

🟩 dm2e:colophon
🟩 dm2e:condition
🟩 dm2e:explicit
🟩 dm2e:incipit
🟩 dm2e:pageDimension
🟩 dm2e:referencedEdition
🟩 dm2e:referencedPublication
🟩 dm2e:relatedPublication
🟩 dm2e:restoration
🟩 dm2e:shelfmark
🟩 dm2e:support
🟩 dm2e:watermark
🟩 dm2e:writtenAreaDimension

EDM for DM2E. Julia Iwanowa, Evelyn Dröge, Steffen Hennicke. Nov 2012

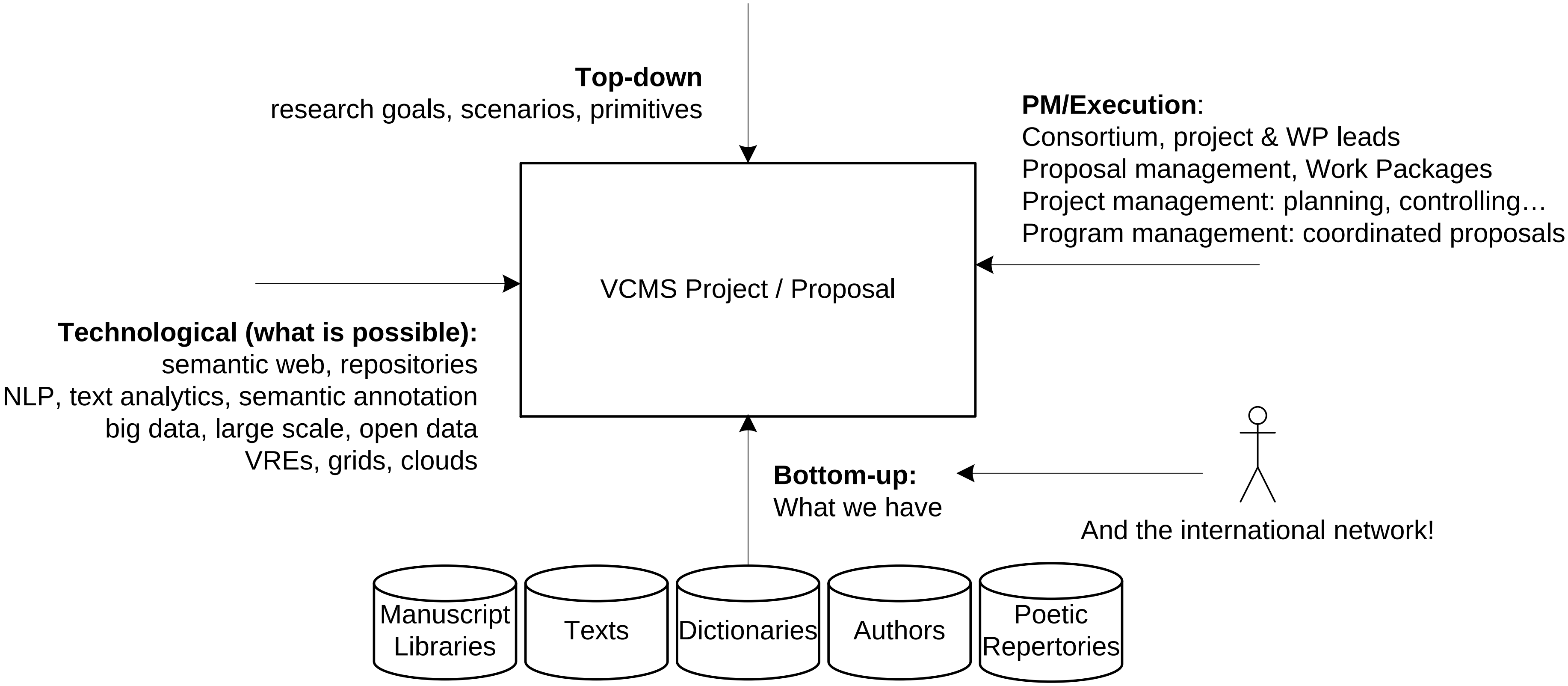
ontotext

VCMS PROJECT DESIGN

I see VCMS as a Program

- Set of interrelated projects having a common goal
- Why
  - You have a very strong foundation: databases & community
  - Integrating the existing databases in a deep way will open new (revolutionary?) avenues of research
  - The full scope is a very ambitious undertaking
  - Start small, expand, build for the future
- Design several proposals according to funding schemes
  - Explore the funding schemes hinted in this presentation. And others!!
  - Select appropriate schemes
  - Design proposals **in compliance** with the schemes
  - Coordinate the sequence of projects
  - **Don't** invest *emotionally* in a proposal: success rate us 15-25%, and there is an element of chance (**do** invest your effort and good thinking)
  - If rejected but the proposal is good, try again (resubmit to another scheme)

VCMS Proposal Inputs



Top-down
research goals, scenarios, primitives

PM/Execution:
Consortium, project & WP leads
Proposal management, Work Packages
Project management: planning, controlling…
Program management: coordinated proposals

VCMS Project / Proposal

Technological (what is possible):
semantic web, repositories
NLP, text analytics, semantic annotation
big data, large scale, open data
VREs, grids, clouds

Bottom-up:
What we have

And the international network!

Manuscript Libraries

Texts

Dictionaries

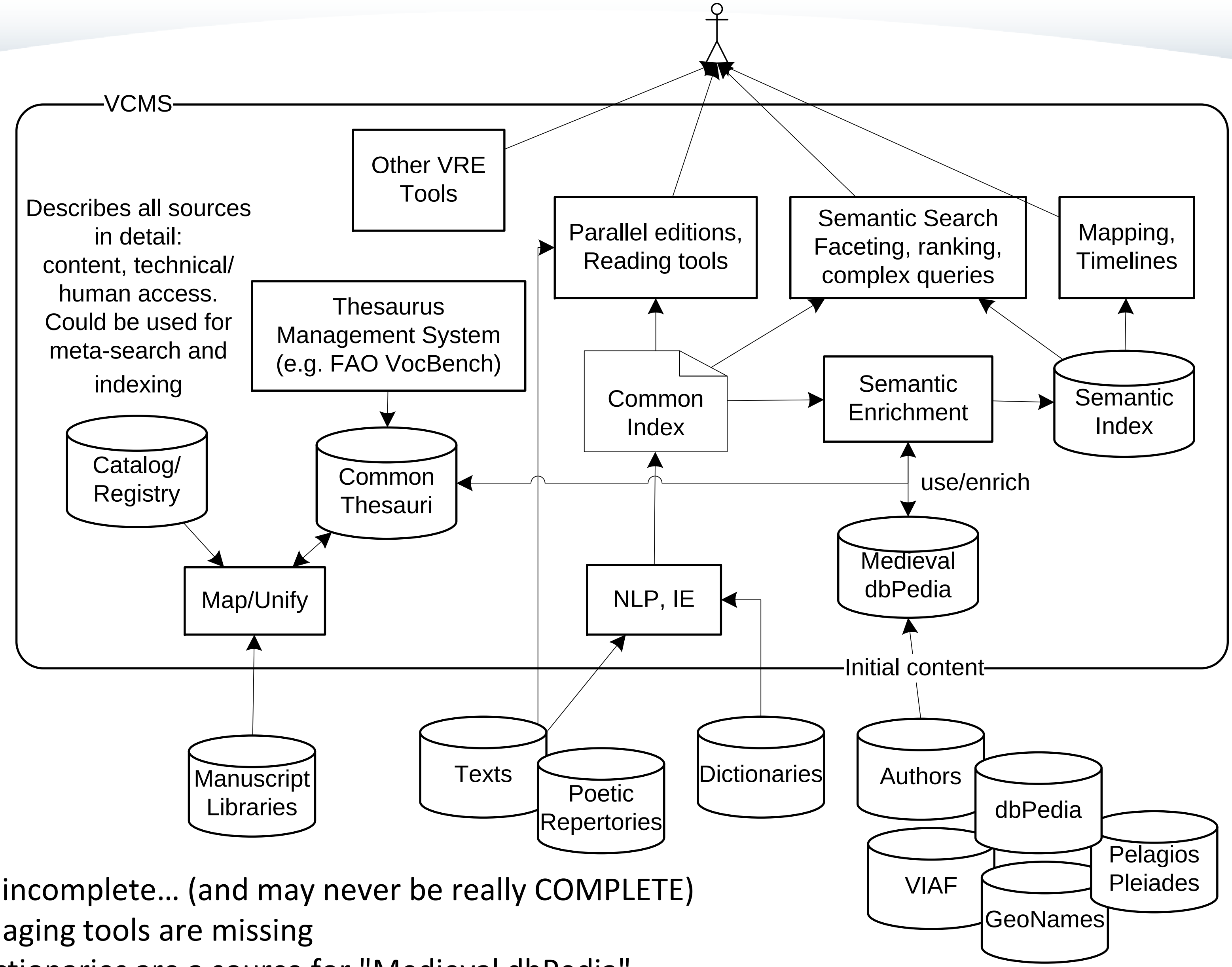
Authors

Poetic Repertories

VCMS Considerations

- Search aggregators (or meta-search: Trame, MegaRep, TraLiRo) are a first IMPORTANT step towards integration
  - In the early days of the web, search aggregators were quite popular (e.g. MetaCrawler) because there weren't very good search engines
  - Now Google (and to a smaller extent Bing and Yahoo) do all you need most of the time, so nobody uses search aggregators
  - Google uses quite a lot of semtech under the hood: Knowledge Graph, schema.org, microdata & microformats, NLP ("do as you mean")
- Shortcomings of search aggregation:
  - Replicates the shortcomings of the original databases
  - Only Union queries, can't do Joins (cross-refer information between two databases)
  - Often a least-common-denominator of individual query languages
  - Can't provide good ranking, which is very important for researchers
- What more can be done by semantic indexing of the databases
  - Powerful full-text index through NLP techniques using the TEI dictionaries
  - Semantic extraction of Entities, Concepts, Relations
  - Information Extraction: document clustering, categorization, classification
  - Unify schemata & terminology across databases (to some extent)
- But who will provide the full text of their database?
  - If you have enough to start with, and do interesting things with them
  - Then others will join! A network/avalanche effect

Possible VCMS Architecture



- This is incomplete… (and may never be really COMPLETE)
- E.g. imaging tools are missing
- E.g. dictionaries are a source for "Medieval dbPedia"

VCMS Conclusions

- You have a lot of data and accumulated knowledge
- Work has been ongoing for 100s of years (>20 years with computers)
- But the IT engagement has been ad-hoc, with small funding, with "do-it-yourself" approaches
- It's an interesting domain: IT specialists will find it fascinating to work on it! Be more daring: *Ask and ye shall receive*
- Many industrial-strength approaches (from Publishing, Life Sciences, etc) can be applied to your domain
- Other CH domains (archaeology, numismatics, linguistics, editions) are already adopting semantic approaches
- Also Consider the Open Data and Open Science initiatives that EU is driving for. **Soon**, scientific contribution won't be measured by publications alone

VCMS Next Steps

At this meeting

- Determine potential VCMS scope (for a first project)
- Appoint people responsible for the proposal
- (maybe) Determine Work Packages and WP leads

After this meeting

- Research and select appropriate funding topics
- Determine overall consortium
- Create proposal structure
- Schedule a proposal writing meeting
- etc

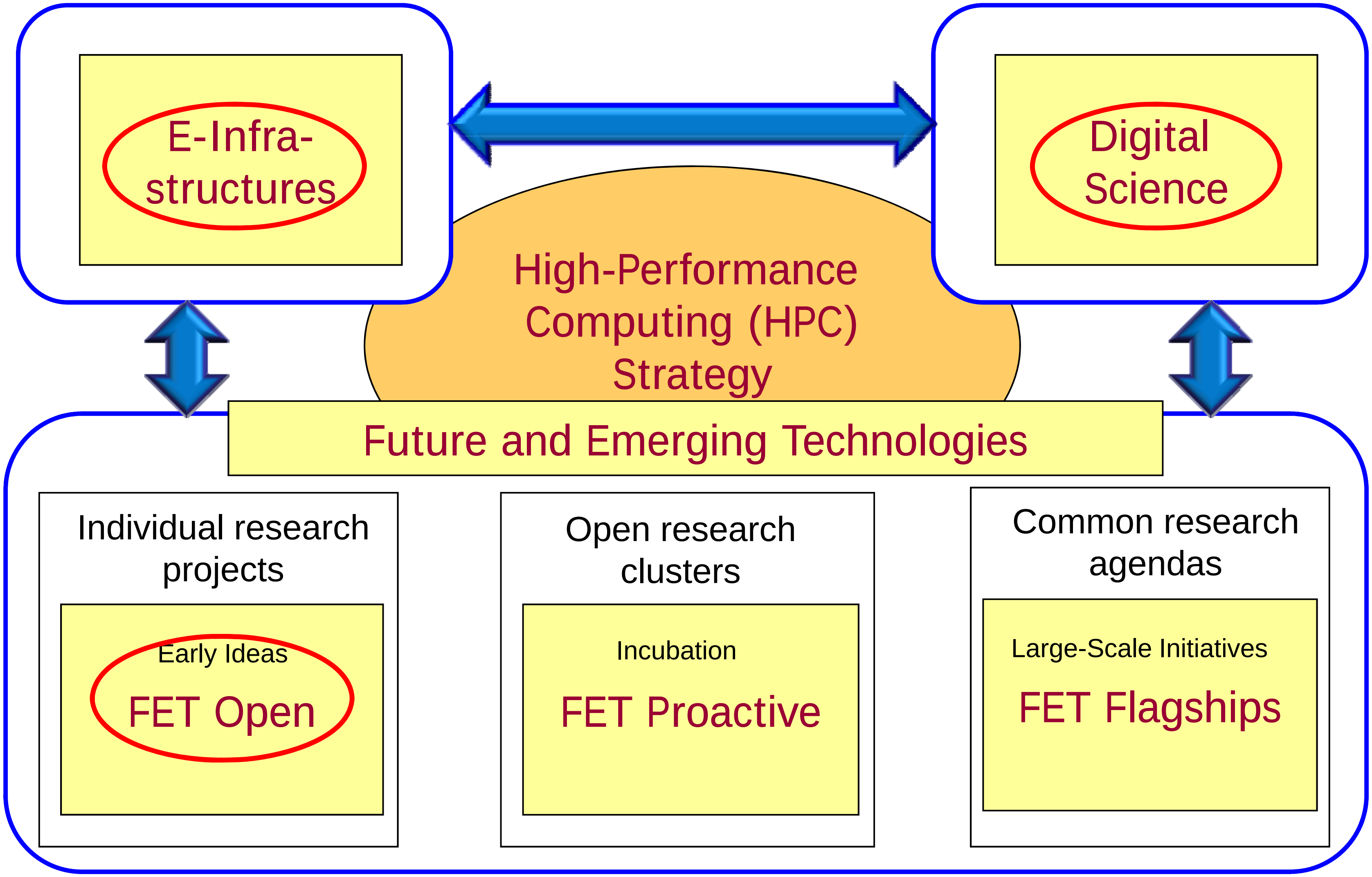
EU H2020 FUNDING INSTRUMENTS

EU Horizon 2020

- H2020 (aka FP8) is EU's science program 2014-2020
  - I'm convinced this is the best way to realize the VCMS envisioned by the COST action, since some participant countries have little access to national funds
- I think the following programs are relevant for us:
  - **FET Open** (Future and Emerging Technologies)
  - **ICT** ("Information and Communication Technologies", part of topic 5 Leadership in enabling and industrial technologies)
  - **eInfrastructures** (shared between DG RTD and DG CONNECT) and **Open Data / Open Science / eScience**
  - **Marie Curie** actions (researcher exchanges, Initial Training Networks, etc)
  - **JRC Frontier research** by the best individual teams (?? not sure)
  - **Social Science and Humanities** (?? only heard about them)
- The info below is from:
  - Presentation "FET in H2020", Roumen Borissov, DG CONNECT (13 Jun 2013)
  - Presentation of Morten Møller, DG CONNECT (5 Sep 2013)
  - Draft of ICT WP 2014–2015 (9 Sep 2013)
- ICT WP to be announced officially at ICT 2013 conference (6-8 Nov 2013, Vilnius)
  - But I don't have enough info about the other programs yet, esp. Humanities

H2020: ICT in Excellent Science

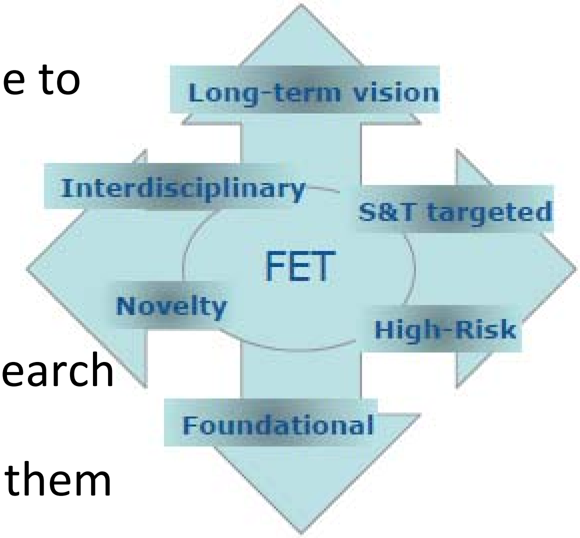
- Schematic on different strands of work. But there are more ICT topics



ontotext

FET Open

- Pros
  - Exploring promising visionary ideas that can contribute to challenges of long term importance for Europe.
  - 'Roots-up' approaches
  - Challenging Current Thinking
  - International Cooperation
  - Stimulates non-conventional targeted exploratory research cutting across all disciplines
  - Exploring and nurturing new research trends, helping them mature in emerging research communities.
  - Short proposals (5-10p), easy to write.
  - Double-blind evaluation, fast decision



- Cons
  - Small projects
  - Low success rate
  - Not an easy path for fast-track innovation

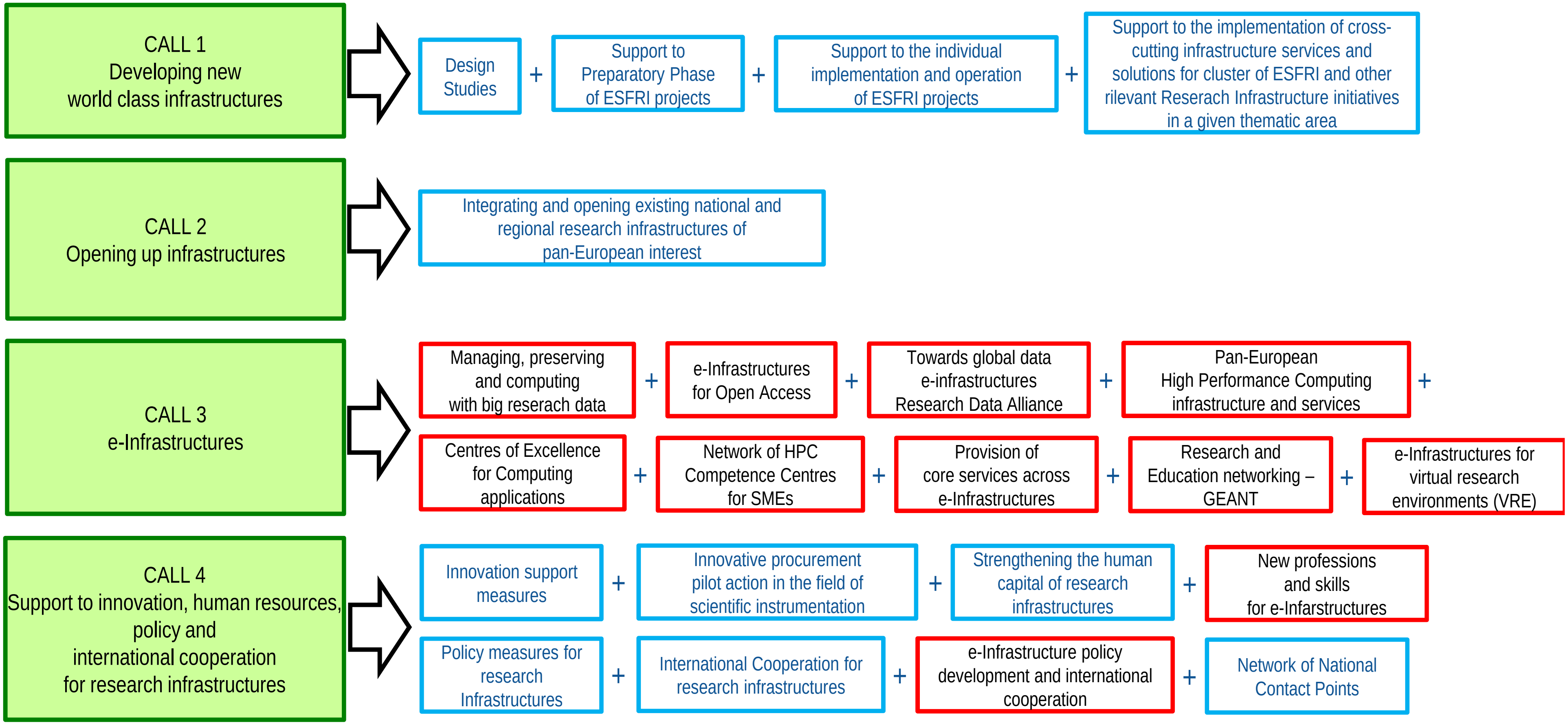
H2020-ICT-2014

- Open Dec 2013, Close 23 Apr 2014, Result Sep 2014
- Potentially applicable topics (code, MEUR):
  - ICT2: 48M Smart System Integration
  - ICT7 :73M Advanced Cloud Infrastructures and Services
  - ICT15: 50M Big data Innovation and take-up
  - ICT17: 15M Cracking the language barrier
  - ICT18: 15M Support the growth of ICT innovative Creative Industries SMEs
  - ICT22: 31M Multimodal and Natural computer interaction
  - ICT30: 7M Human-centric Digital Age
- Need to research and decide which are applicable!

H2020-ICT-2015

- Open Jul 2014, Close 20 Jan 2015, Result Jun 2015
  - ICT8 : 22M Boosting public sector productivity and innovation through cloud computing services
  - ICT16: 39M Big data - research
  - ICT19 :41M Technologies for creative industries, social media and convergence.
  - ICT20: 52M Technologies for better human learning and teaching

Research Infrastructures WP 2014-2015



CALL 1
Developing new
world class infrastructures

Design Studies + Support to Preparatory Phase of ESFRI projects + Support to the individual implementation and operation of ESFRI projects + Support to the implementation of cross-cutting infrastructure services and solutions for cluster of ESFRI and other rilevant Reserach Infrastructure initiatives in a given thematic area

CALL 2
Opening up infrastructures

Integrating and opening existing national and regional research infrastructures of pan-European interest

CALL 3
e-Infrastructures

Managing, preserving and computing with big reserach data + e-Infrastructures for Open Access + Towards global data e-infrastructures Research Data Alliance + Pan-European High Performance Computing infrastructure and services +

Centres of Excellence for Computing applications + Network of HPC Competence Centres for SMEs + Provision of core services across e-Infrastructures + Research and Education networking – GEANT + e-Infrastructures for virtual research environments (VRE)

CALL 4
Support to innovation, human resources, policy and
international cooperation
for research infrastructures

Innovation support measures + Innovative procurement pilot action in the field of scientific instrumentation + Strengthening the human capital of research infrastructures + New professions and skills for e-Infarstructures

Policy measures for research Infrastructures + International Cooperation for research infrastructures + e-Infrastructure policy development and international cooperation + Network of National Contact Points

- RI Call 1: Open Oct 2014. Close maybe Feb-Mar 2015?

Discussion

- Thanks for listening!
- vladimir.alexiev@ontotext.com