VU Digital History Datathon
2 April from 9.15-18.00
(Room C161 W&N Building VU Amsterdam)

Idea
Spend one day linking a bunch of (VU-produced) E-historical RDF datasets together and show what cool stuff you can do with this dataset. Also, Pizza.

Programme:

9.00-9.30: Get together
9.30-9.35: Goals and setup of the datathon
9.35-10.15: Data pitches by participants
(short presentation on linking data: http://www.slideshare.net/vdeboer/linking-up-your-data)
10.15-16.30: Hacking data
16.30-17.30: Present results
17.30-18.00: Closing

Participants:

- Chris Dijkshoorn, Albert Merono, Victor de Boer, Niels Ockeloen (organization)
- Ivo Zandhuis (gemeentegeschiedenis.nl) + 2
- ~~Stevan Rudinac (via Lora)~~
- ~~Niels Groot-Obbink (via Lora)~~
- Silvia
- Valentina
- Wouter Beek
- Thijs van Beek

Visitors:

- Serge ter Braake

Core Datasets:

- Niels: (sample set of) BiographyNet
- Albert: STCN, CEDAR data (historical censuses)
- Chris: Rijksmuseum Collection, maybe parts of the Naturalis collection (animals)
- Victor: Verrijkt Koninkrijk and Dutch Ships and Sailors

- Ivo Zandhuis: historical municipalities dataset (gemeentegeschiedenis.nl, histopo.nl, api.histopo.nl)
- Thijs: IconClass & BibleOntology

Other data that might be interesting:

- Amsterdam Museum (VU)
- Agora (VU)
- Stadsarchief
- War monuments (4&5mei) (VU)
- Rijksmonumenten

Possible Goals:

- Victor: link DSS/VK to other datasets and/or make a small web application showcasing the benefits of LOD here.
- Victor: make a sustainable VU e-human datacloud
  - for reuse by students/researchers
  - Draw cloud
  - Formulate federated sparql query
- Victor: Write submission for ISWC (idea from Paul) http://iswc2014.semanticweb.org/call-replication-benchmark-data-papers
  - Make cool links and apps first, worry about paper later
- Chris: Link data to artworks depicting naval battles of the Rijksmuseum, e.g.: https://www.rijksmuseum.nl/nl/collectie/SK-A-22/zeeslagen https://www.rijksmuseum.nl/nl/zoeken?p=1&ps=12&f.classification.iconClassIdentifier.sort=45H3(%2B3)&ii=0
- Chris: Linking Iconclass branch to CEDAR (e.g. http://www.iconclass.org/rkd/4/)
- Niels: Investigate what would be interesting links to make between the BiographyNet data set and the other participating data sets.
- Albert: Show added value of linking these (primarily CEDAR) to other datasets to answer historical questions
- Albert: Find a straightforward way of standardizing census variables
- Thijs: Find matches between IconClass (Code 7) and BibleOntology
  - http://iconclass.org/rkd/0/
  - http://bibleontology.com/home/

Dataset descriptions

STCN (Short Title Catalogue of the Netherlands)
SPARQL endpoint: http://lod.cedar-project.nl:8080/sparql/cedar
Raw dataset: http://lod.cedar-project.nl/~amp/stcn/stcn-dump.nt.gz
http://lod.cedar-project.nl/~amp/stcn/knuttel.ttl
http://lod.cedar-project.nl/~amp/stcn/weekhout.ttl
http://lod.cedar-project.nl/~amp/stcn/out-knuttel.ttl

http://lod.cedar-project.nl/~amp/stcn/out-weekhout.ttl
Dereferenceable URIs: http://stcn.data2semantics.org/
Data description: The stcn-dump.nt.gz contains the raw database with info about publications, their authors, publishers, genera, etc. Example SPARQL queries are available here. Additionally, knuttel.ttl and weekhout.ttl are similar datasets with lists of prohibited / banned books, and out-knuttel.ttl and out-weekhout.ttl contain mappings between these prohibited books and their STCN equivalents.

CEDAR (Dutch historical censuses 1795-1971)
SPARQL endpoint: http://lod.cedar-project.nl:8080/sparql/stcn
Raw dataset: https://github.com/CEDAR-project/DataDump
Dereferenceable URIs: N/A
Data description: The dataset contains 507 named graphs (listed here) that can be queried following the RDF Data Cube data model (examples here). To query all 507 named graphs at the same time, the graph group http://lod.cedar-project.nl/resource/cedar-dataset can be used in the FORM clause.
Other graph groups:
TabLinker translation: http://lod.cedar-project.nl/resource/v1/cedar-dataset

Half-curated translation: http://lod.cedar-project.nl/resource/v2/cedar-dataset

Harmonized release: http://lod.cedar-project.nl/resource/r1/cedar-dataset

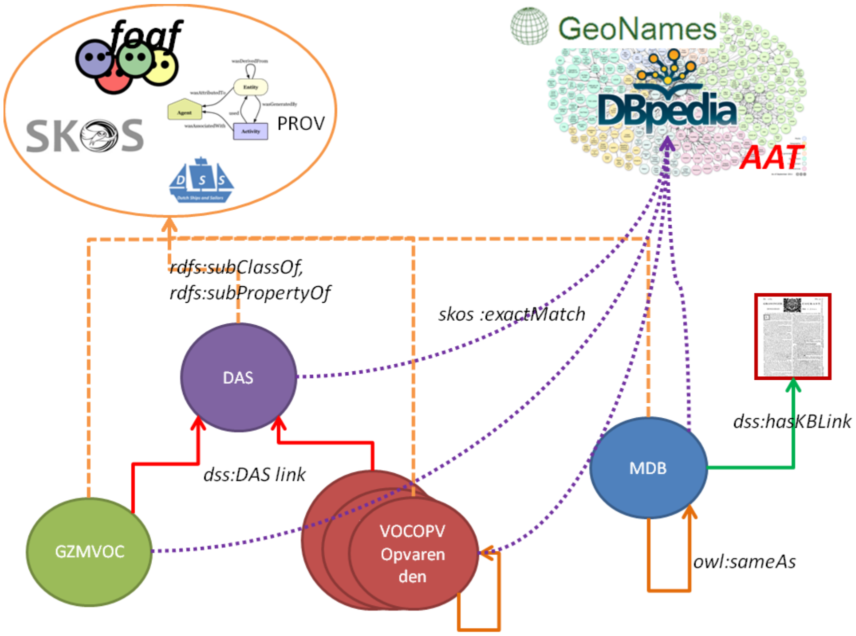
Dutch Ships and Sailors
SPARQL endpoint: http://semanticweb.cs.vu.nl/dss/sparql/
ClioPatria interface: http://semanticweb.cs.vu.nl/dss/sparql/
Data at github: https://github.com/biktorrr/dss
Dereferenceable URIs: yes, for example
http://purl.org/collections/nl/dss/vocopv/opvarenden-344716
Dataset description:These are actually four interconnected datasets about Dutch maritime history, created in the context of the CLARIN DSS project. Specifically they are:

1. The Monsterrollen databases, which contain elaborate data on the crew composition of shipsfrom the Northern Netherlands (c. 1800-1930) and provide information on the sailors involved,such as the places of origin, wage and age. (*namespace http://purl.org/collections/nl/dss/mdb*) This dataset has a number of sameAs identifying ships as well as links to KB historical newspapers.
2. The VOC Opvarenden database, providing extensive data on crews of VOC ships leaving the Dutch Republic.(*http://purl.org/collections/nl/dss/vocopv/)*
3. The Dutch-Asiatic Shipping (DAS) database, providing data on all inter-continental voyages of VOC ships. (*http://purl.org/collections/nl/dss/das*)
4. The Generale Zeemonsterrollen (GZM) database, providing data on the crew composition and sometimes location of VOC ships stationed in Asia and not engaged in

inter-continental shipping.(*http://purl.org/collections/nl/dss/gzmvoc*)

Links (see image)

- The three latter datasets have inter-dataset links based on DAS identifiers.
- All datasets' schemas are mapped to a central DSS schema, which in turn is mapped to FOAF, SKOS etc.
- Links to AAT (shiptypes, ranks) exist
- Links to GeoNames (places in three datasets) exist
- A few links to DBPedia exist
- MDB has links to KB historical newspapers



Verrijkt Koninkrijk

SPARQL endpoint: http://semanticweb.cs.vu.nl/verrijktkoninkrijk/sparql/

Dereferenceable URIs: yes for example: http://purl.org/collections/nl/niod/entity-MUSSERT

Data dumps: https://github.com/biktorrr/Verrijktkoninkrijk

Dataset description: The Verrijkt Koninkrijk Data concerns Dr Loe de Jong's Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog and was based on the PDFs as provided by NIOD at http://www.niod.knaw.nl/koninkrijk/ . The books have been OCRed and transformed to structured XML by researchers from the Universiteit van Amsterdam. This data is available through www.loedejongdigitaal.nl. A search interface is available at http://search.loedejongdigitaal.nl. A resolver server was installed which responds by presenting the structure (in XML) when presented with a URL. For example, http://resolver.loedejongdigitaal.nl/nl.vk.d.1.6.1.43 is resolved to the XML fragment of that paragraph. Removing the last number of the identifier (43) results in its broader section, etcetera. Paragraphs are the smallest logical units (also, a page is not a logical unit). We provide two RDF 'stepping stones' into the book text. The 'Back of the Book index' and the 'Named Entities index'. Both are SKOS vocabularies and consist of terms pointing to resolver.loedejongdigitaal.nl URIs. These vocabularies are linked to external sources as well. All RDF is available as Linked Data at the VK Semantic Layer at http://semanticweb.cs.vu.nl/verrijktkoninkrijk/ The base namespace for the VK/NIOD triples is http://purl.org/collections/nl/niod/ (abbreviated as niod:). Datasets, mapping sets and schemata are all loaded as separate named graphs (http://semanticweb.cs.vu.nl/verrijktkoninkrijk/browse/list_graphs).

BiographyNet
ClioPatria interface: http://eculture2.cs.vu.nl:1349
SPARQL Endpoint: http://eculture2.cs.vu.nl:1349/sparql/
Data dump: no
Dereferenceable URIs: not yet (see below)
Dataset description: The Biography Portal of the Netherlands links more than 75.000 Dutch people mentioned in various databases, through a limited set of metadata. This project aims to enhance its potential for historical research by transforming the available data into a semantic knowledge base and through the creation of a demonstrator. BiographyNet is a multidisciplinary project that combines expertise from history, computer science and computational linguistics. An interlinked semantic knowledge base will be created by extracting relations between people, places, historic events and time periods from biographical descriptions.

Unfortunately there is no consensus yet on what would be the ideal identifier for individual biographies. This has to do with the fact that these biographies come from different sources and use different identification systems. On top of that, the Biography Portal of the Netherlands uses its own identifier system. To add to the confusion, the XML serialisation used a random indexing system for multiple biographies for the same person. Hence; the change of these IDs changing in the near future makes linking a bit cumbersome, but we want to investigate which (types of) links would be interesting and mine new contacts for this and ideas.

Rijksmuseum Collection
SPARQL endpoint: http://semanticweb.cs.vu.nl/clustersearch/sparql/
Dereferenceable URIs: no
Data dumps: should check license wise
Dataset description: The Rijksmuseum collection consists of around 1.000.000 artworks, including works by Dutch masters like Rembrandt and Vermeer. 600.000 of these artworks belong to the print collection. Only a fraction of the collection is available as digitised representations, around 180.000 prints and 70.000 others have digital representations.

After a picture is taken of the artwork in a dedicated studio, catalogers add metadata to the records. When possible annotations are used originating from structured vocabularies. For this purpose a general in-house thesaurus was created, containing concepts describing materials, techniques, locations, temporal concepts and events. In total there are 65.000 concepts.

Additionally, there is a dedicated thesaurus containing information about people, containing information about over 77.000 persons.

gemeentegeschiedenis.nl

SPARQL endpoint: none
Raw dataset: (i.e. downloadable RDF dumps) per province, eg:
http://www.gemeentegeschiedenis.nl/provincie/rdfxml/Flevoland Not all data is available here
Dereferenceable URIs: through our domain name; Just use our URIs, like
http://www.gemeentegeschiedenis.nl/gemeentenaam/Amsterdam (with content negotiation) or use the 'directory' rdfxml to address te rdf without content negotiation:
http://www.gemeentegeschiedenis.nl/gemeentenaam/rdfxml/Amsterdam
Data description: We have given each (former) municipality its own uri. In the rdf you'll find the name(s) for the municipality, its id's in different coding systems (Amsterdamse Code and CBS-code), links to wikipedia and dbpedia and start- and end dates.

Other datasets:

Other data that might be interesting:

- Amsterdam Museum (VU): http://semanticweb.cs.vu.nl/europeana/sparql/
- Agora (VU)
- Stadsarchief
- War monuments (4&5mei) (VU) http://semanticweb.cs.vu.nl/verrijktkoninkrijk/sparql/
- War monuments (4&5mei) (VU) http://semanticweb.cs.vu.nl/verrijktkoninkrijk/
- Rijksmonumenten
- BBC enrichment: http://eculture2.cs.vu.nl:3020/sparql/ (SPARQL query for retrieving programs: SELECT DISTINCT ?prog WHERE {?prog <http://data.vista-tv.eu/ontologies/epg/enrichmenthas_synopsis_enrichment> ?b .})

Linking Occupations

Linking HISCO and ICONCLASS doc:
https://docs.google.com/spreadsheets/d/1fEToVGoMJSz580h1xqktOBm0s3RpRh6zos-L2EgWupA/edit#gid=0

HISCO search interface: http://historyofwork.iisg.nl/search.php
Hisco namespace: http://historyofwork.iisg.nl/list_micro.php?keywords=

CEDAR-HISCO namespace: @prefix cedar: <http://cedar.example.org/ns#>
<http://cedar.example.org/ns#hisco-39310>
<http://cedar.example.org/ns#hisco-87340>

ICONCLASS http://iconclass.org/rkd/47L2/

Pizzas

Add your name and name your favourite pizza from http://www.thuisbezorgd.nl/palinuro

Albert: quattro stagioni
Victor: pizza nutella
Chris: pizza carpaccio
Menno: **Caprese**
Silvia: Quattro stagioni
Myriam: Margherita
Niels: Margherita

LINK RESULTS (whoo)

- Niels: links-RMA-naar-BNET.trig
  - 5000+ links from people in the BiographyNet RDF data to people in the Rijksmuseum RDF data. Automatic generation of skos:closeMatch links between people in BiographyNet and the Rijksmuseum Amsterdam collection data sets, using a SPARQL query in ClioPatria;
  - PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> PREFIX bnet: <http://purl.org/collections/nl/biographyned/> PREFIX rma: <http://purl.org/collections/nl/rma/schema#> PREFIX rmap: <http://purl.org/collections/nl/rma/people/> PREFIX skos: <http://www.w3.org/2004/02/skos/core#> CONSTRUCT { ?bioPerson skos:exactMatch ?rmaPerson } WHERE {   ?rmaPerson rma:name ?commonName.   ?rmaPerson rdf:type rma:Person.   ?bioPersDes bnet:persName ?commonName.   ?bioBioDes bnet:person ?bioPersDes. ?bioBioDes bnet:aggregatedPerson ?bioPerson. }
  - Use this SPARQL query to generate the skos:exactMatch links between people in BiographyNet and the Rijksmuseum Amsterdam collection data sets.
  - Using <http://users.ugent.be/~tdenies/up/>, contentConfidence was determined ad 0.5 (50/50), as these links were made using string matching on names only. Other properties such as birth- and/or dead dates could be used to increase confidence, however within the time limitations of this event further improvement was not carried out. Nonetheless, this event really showed the great potential for linking the BiographyNet database to other data sources, which is one of the main outcomes of this event for BiographyNet.
- Victor: Victor_datathon_links.ttl:

    - 2 links from DSS to RM,
        - One print of a ship (POLLUX)
        - One print of a ship type
    - 61 links from DSS Ranks to CEDAR Hisco URIs
        - These link ship ranks (captain, boatsman etc) to hisco codes. This means we can now ask for a ship and its captain what the total number of captains in the city of birth was.
- Albert
    - 1320 links of CEDAR municipalities (by Amsterdamse Code) to gemeentegeschiedenis.nl municipalities (same Amsterdamse Code) in file **cedar2gg-links.ttl**
    - 33 links of ICONCLASS to HISCO occupations, reading Chris/Myriam links from this Google Spreadsheet using this quick and dirty script, in file **gspread-ic-hisco.ttl**

The paper

ISWC datasets call

http://iswc2014.semanticweb.org/call-replication-benchmark-data-software-papers

Data introduces an important data set to the community. This highly important task is often difficult to publish, as its main contribution lies in providing others the means for accomplishing their goals. Even though dbpedia and wordnet are some of the most valuable and widely used resources in our community, and have made an invaluable contribution to our science, they were very difficult to publish as papers. For example:

    - S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak, Z Ives. Dbpedia: A nucleus for a web of open data. ISWC 2007. [PDF]

Review Criteria: Is there a similar data source? Is the source of interest to the semantic web community (and society in general)? Is the source semantic, linked, etc.? Does it use URIs. Is it available to the community? Was the data used for something scientific, practical, etc.? Is the data likely to be repurposed for other uses?