Linked Open Data and Semantic Technologies for Research in Agriculture and Forestry

Platform Linked Data Nederland – 2 April 2015
Rob Lokers, Alterra, Wageningen UR
Contents

- Data related challenges in agricultural (and forestry) research
- Exploiting Linked Data and semantic technologies
  - Trees4Future: a Research Infrastructure for forestry research
  - SemaGrow: a LOD infrastructure supporting the agricultural community
  - some examples and lessons learned
Challenges in Agricultural & Forestry Research

- Research data is only partially available for the whole (research) community
- Data is:
  - stored locally/privately, in silos
  - not accessible
  - not documented and metadata is not generated
- No incentive nor sense of urgency to actively / automatically share data other than through networks and personal contacts.
- Thus, valuable research data is hard to find if you don’t know the right people

- However, increased pressure to document data, publish research results as open data in a comprehensible and usable manner!
Challenges in Agricultural & Forestry Research

- Agricultural & forestry researchers require data from different domains and has usually very detailed specifications
  - example domain meteorology: Many ways exist to measure, predict, aggregate, post-process temperature or precipitation data. You need quite some technical expertise on climate to be able to select the most appropriate data for your job.

- Required data is stored at different locations, documented by different institutions working in different domains with different objectives and at different “quality levels”

- Metadata often does not provide or easily reveal the relevant details, does not provide the required depth and structure and does not reveal characteristics and patterns of the data itself
However, most of the complexity we are struggling with is caused above all by structural insufficiencies due to the networked nature of our society. The specialist nature of many enterprises and experts is not yet mirrored well enough in the way we manage information and communicate. Instead of being findable and linked to other data, much information is still hidden.

With its clear focus on high-quality metadata management, Linked Data is key to overcoming this problem. The value of data increases each time it is being re-used and linked to another resource. Re-usage can only be triggered by providing information about the available information. In order to undertake this task in a sustainable manner, information must be recognised as an important resource that should be managed just like any other.

Linked Open Data: The Essentials A Quick Start Guide for Decision Makers, Florian Bauer & Martin Kaltenböck
Linked Open Data in agricultural research

Research projects: Trees4Future, SemaGrow

- Trees4Future: EU (Forestry) Research Infrastructure project
- SemaGrow: EU ICT Research project
- Closing the gap between data supply and data demand in agricultural & forestry research
  - improving data availability, harmonization, discoverability
  - supporting researchers and research processes (data processing & data analytics)
  - using Linked Open Data and semantic technologies
Trees4Future – Research Infrastructure

- Setting up a European knowledge network
- Explaining the benefits of data sharing
- Organizing activities to collect, structure and harmonize forestry data
- Setting up a “Clearinghouse” as an operational forestry metadata repository
  - making European datasets discoverable and accessible for the whole community.
  - using open standards to register and harvest metadata into a centralized metadata repository
  - using open standards to access data (services, datasets)
  - using LOD and semantic technologies to improve discoverability of datasets
Discover forestry research datasets

Keyword
(dan three letters)

<table>
<thead>
<tr>
<th>Where?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southwest</td>
</tr>
<tr>
<td>Norheast</td>
</tr>
</tbody>
</table>

| Exact | Contains | Starts with |

<table>
<thead>
<tr>
<th>When?</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
</tr>
<tr>
<td>To date</td>
</tr>
</tbody>
</table>

Search for keyword: diameter; leaf coloration

Results: 15 total

<table>
<thead>
<tr>
<th>#</th>
<th>Name (click to see metadata)</th>
<th>Description</th>
<th>Publisher</th>
<th>Score</th>
<th>Services</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F03 x S103</td>
<td>Crossing; Quercus robur</td>
<td>Austrian Institute of Technology - AIT; Stephan Gaußlicher; <a href="mailto:stephan.gaubitser@ait.ac.at">stephan.gaubitser@ait.ac.at</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>L10Q</td>
<td>Association Population; Quercus petraea</td>
<td>INRA; Véronique Jorge; <a href="mailto:jorge@orleans.inra.fr">jorge@orleans.inra.fr</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>L12Q</td>
<td>Association Population; Quercus petraea</td>
<td>INRA; Véronique Jorge; <a href="mailto:jorge@orleans.inra.fr">jorge@orleans.inra.fr</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>L14Q</td>
<td>Association Population; Quercus petraea</td>
<td>INRA; Véronique Jorge; <a href="mailto:jorge@orleans.inra.fr">jorge@orleans.inra.fr</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>L16Q</td>
<td>Association Population; Quercus petraea</td>
<td>INRA; Véronique Jorge; <a href="mailto:jorge@orleans.inra.fr">jorge@orleans.inra.fr</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>L1Q</td>
<td>Association Population; Quercus petraea</td>
<td>INRA; Véronique Jorge; <a href="mailto:jorge@orleans.inra.fr">jorge@orleans.inra.fr</a></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Find rainfall data

Also look for data on precipitation, rain...

Collects metadata of distributed datasets

Using ontologies: link precipitation to rainfall

External ontologies

Describes the datasets e.g. keyword = precipitation

Wageningen UR
For quality of life
Example – semantic tagging & search
Climate dataset - metadata contains term rain

Trees4Future

- T4F dataset
- T4F Concept 287
- Clearinghouse semantic tagging
  - skos:related
  - skos:prefLabel
  - skos:altLabel

AGROVOC

- Rainfall
- Rain
- AgroVoc concept 6435
  - skos:exactMatch
  - skos:narrower
- AgroVoc concept 6156
  - skos:narrower
- AgroVoc concept 2360
  - skos:altLabel
  - climate
- Snow
  - skos:altLabel
- Precipitation
  - skos:altLabel

Clearinghouse harvester analyses metadata (NLP) and links dataset with T4F concept.

External ontology knows concept rain.
Some lessons learned

- Metadata is not always perfect and unambiguous
  - limited or no metadata available for (research) datasets
  - different spelling, use of abbreviations etc.
  - use of metadata fields by editors is not consistent
  - link to vocabularies often absent
  - metadata supplied by “non-experts”, post-project

- Automatic semantic tagging (using NLP) is not an easy job
  - available ontologies are often very specialized and/or not complete
  - lot of potential ambiguity
  - most application in this area are on bibliographic information, where in general much more “context” is available to work with.

- Querying and reasoning over semantic network is a challenge
  - performance issues require “undesirable” optimizations
  - no generic recipe to determine “relevance”

- Awareness is growing
SemaGrow - Objectives

Problem statement:

*LOD network is growing, data gets interconnected but it is still not easy to transparently access this distributed cloud of heterogeneous data sources.*

- Extend LOD capabilities by setting up an infrastructure that:
  - Allows transparent, federated access to heterogeneous distributed (big) data sources through one federated (SPARQL) endpoint
  - is efficient, real-time responsive, and scalable
  - Is flexible and robust enough to allow data providers to publish in the manner and form that best suites their purposes, and data consumers to query in the manner and form that best suits theirs.
SemaGrow - Objectives

- Test and evaluate the infrastructure through implementation and evaluation of (agricultural) use cases
  - develop agricultural use cases
  - design & implement demonstrators
  - test & evaluate performance

SemaGrow application show cases

- FAO – Information Management (Agris, AGROVOC)
- Alterra, Wageningen UR – Agricultural & Forestry Modelling
- AgroKnow – Agricultural Education
SemaGrow - Use case agricultural research

An Example:

Kenneth is an agricultural modeller in Kenya

- wants to assess consequences of climate change on agricultural yields
- needs input for his models:
  - temperature, precipitation
  - soil (available)
  - crop trial data
- knows about AgMIP, a global community on agricultural modelling
- knows that Joe is active in AgMIP
SemaGrow- Use case agricultural research

- Can data analytics and data processing be improved by describing not only the semantics of the datasets but also of the contained data?
- Can we build an infrastructure that supports semantic querying of big linked datasets?
- Can these improvements be integrated in existing applications to better support research data requirements?

Example query:
- Identify available crop experimental data for the Mediterranean area where the crop is sunflower and the soil type is sandy soil.
- Data sources: various crop trial databases, European soil map
- Evaluates the system’s ability to perform semantic searches over the dataset metadata, by matching “Mediterranean” with crop trial spatial characteristics and “sandy soil” with crop trial soil characteristics.
## SemaGrow data sizes

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Current (Jan. 2014) Data Size (GTriples)</th>
<th>Average Growth Rate (% per year)</th>
<th>Projected Data Size (end of 2015)</th>
<th>Projected Data Size (end of 2020)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Heterogeneous Data Collections &amp; Streams</td>
<td>12,986.69</td>
<td>25.84%</td>
<td>25,877.67</td>
<td>1,375,876.80</td>
</tr>
<tr>
<td>(B) Reactive Data Analysis</td>
<td>10,385.07</td>
<td>12.93%</td>
<td>14,957.83</td>
<td>809,846.85</td>
</tr>
<tr>
<td>(C) Reactive Resource Discovery</td>
<td>1,155.29</td>
<td>8.24%</td>
<td>1,465.17</td>
<td>2,183.08</td>
</tr>
<tr>
<td>TOTAL (A+B+C)</td>
<td>24,527.05</td>
<td>19.73%</td>
<td>42,300.67</td>
<td>2,187,906.73</td>
</tr>
</tbody>
</table>

* Tera Scale

* Peta Scale
SemaGrow architecture
SemaGrow architecture

- **Resource Discovery**
  - Query patterns
  - Equivalent patterns
  - Candidate sources

- **Query Transformer**
  - Query patterns
  - Equivalent patterns

- **Query Decomposition**
  - Measurements
  - Query strategy

- **Federated Query Manager**
  - Query
  - Results
  - Query
  - Results

- **Data Consumer**
  - Query
  - Client Parameters
  - Query results

- **Data provider**
  - Query
  - Results

- Contents and schema metadata
SemaGrow architecture

What’s the best strategy for the requested query?

Data Consumer
- Query
- Client Parameters

Query Decomposition
- Query strategy
- Measurements
- Query patterns
equivalent patterns

Federated Query Manager
- query
- results
- query
- results

Data provider
- query
- results

Data Consumer
- Query
- Client Parameters

Query Transformer
- query patterns
equivalent patterns

Resource Discovery
- query patterns
candidate sources

contents and schema metadata
Where should I look? How many results will I get and how fast?
SemaGrow architecture

- **Data Consumer**
  - Query
  - Client Parameters
  - Query results

- **Query Decomposition**
  - Query strategy
  - Query results

- **Federated Query Manager**
  - Query
  - Results
  - Query
  - Results

- **Resource Discovery**
  - Query patterns
  - Candidate sources

- **Ontology alignment**
  - Query Transformer
  - Equivalent patterns

- **Data provider**
  - Contents and schema metadata

**Questions:**
- Are there equivalent data in an aligned schema?
What’s the best strategy for the requested query?
SemaGrow architecture

Resource Discovery

Query Transformer

Query Decomposition

Federated Query Manager

Data Consumer

Data

provider

Data

provider

Query strategy

Client Parameters

Query results

Is my strategy working? Can I improve overall performance?
SemaGrow architecture

Resource Discovery

Query Transformer

Query Decomposition

Federated Query Manager

Data Consumer

Data provider

Data provider

query patterns
equivalent patterns
measurements
query strategy
query
results
query
results
query patterns
candidate sources
contents and schema metadata

Client Parameters

Query

Query results
Some lessons learned (up till now...)

- Many technical pitfalls exist...
- Building on rather immature technology (e.g. RDF databases) and semantic networks
- Federated access could work, but:
  - many potential “points of failure”
  - ontology alignment is complicated, even in one domain
- Big Linked Data is an enormous challenge, maybe not realistic?
  - querying RDF data structures does not perform well yet
  - even with all kinds of optimizations (which are sometimes against the LOD principles)
  - will we ever really triplify Gbyte datasets?

On the other hand:
- This is ICT research...
- Even if small steps can be made, there can be high benefits!
Thanks for your attention!