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

Platform Linked Data Nederland – 2 April 2015

Rob Lokers, Alterra, Wageningen UR

WAGENINGEN UR
For quality of life

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 a 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

Challenges in Agricultural & Forestry Research

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

TREES 4 FUTURE

HOME ABOUT TRANSNATIONAL ACCESSES PUBLICATIONS NEWS EVENTS CONTACT LINKS DATABASE ACCESS INTRANET

Trees4Future in brief

Theme: Research infrastructures for forestry research
Duration: 4 years
Budget: app. 9 Mill EUR
Funder: EU 7th Framework Programme (FP7)
Partners: 28 organisations

Download the T4F brochure (pdf)

Discover forestry research datasets

Keyword (min three letters): diameter; leaf coloration;
Exact Contains Starts with

Where?
Southwest
Northeast
Overlap Within

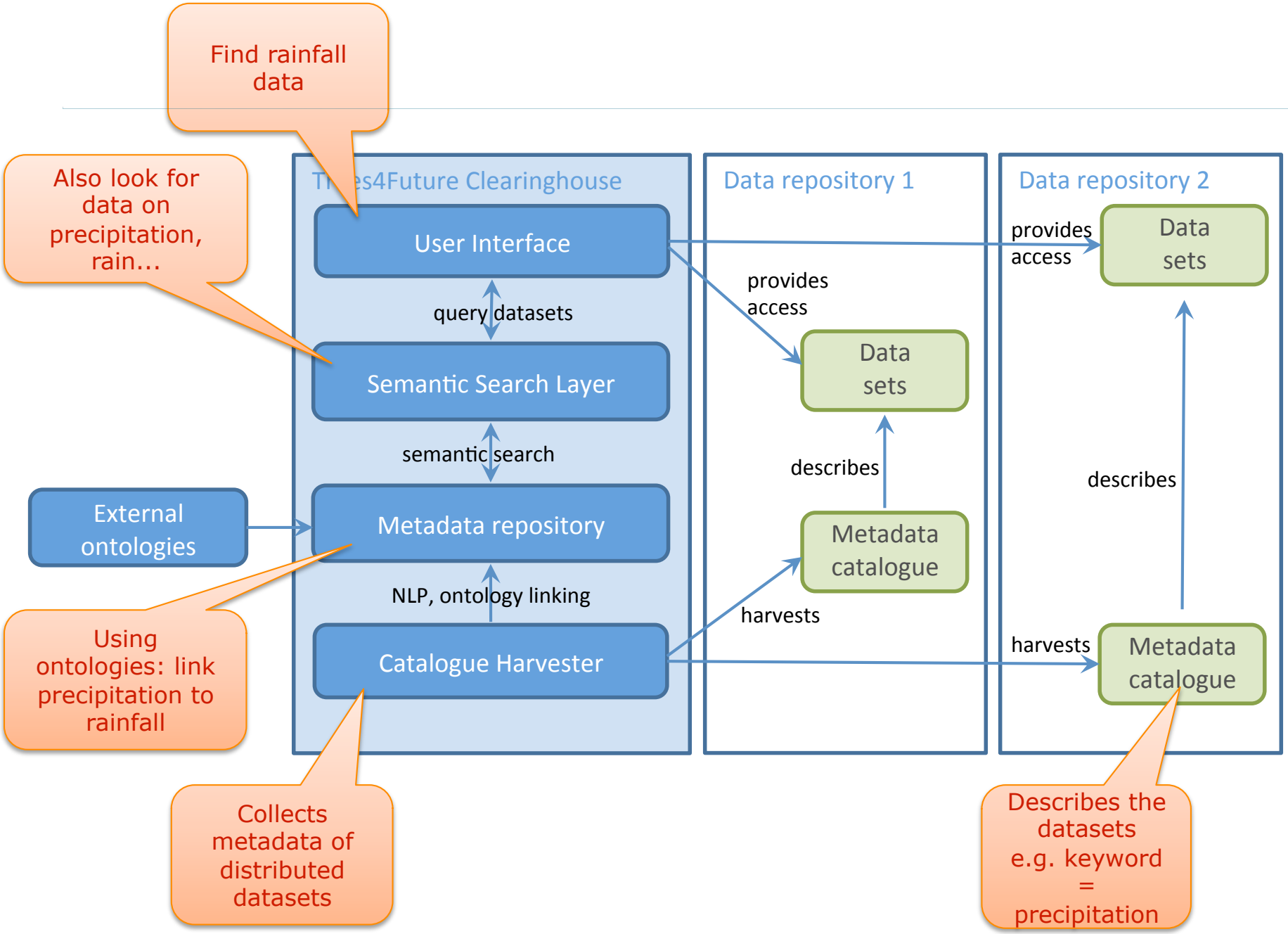
When?
From
To date

Search

Search for keyword: diameter; leaf coloration

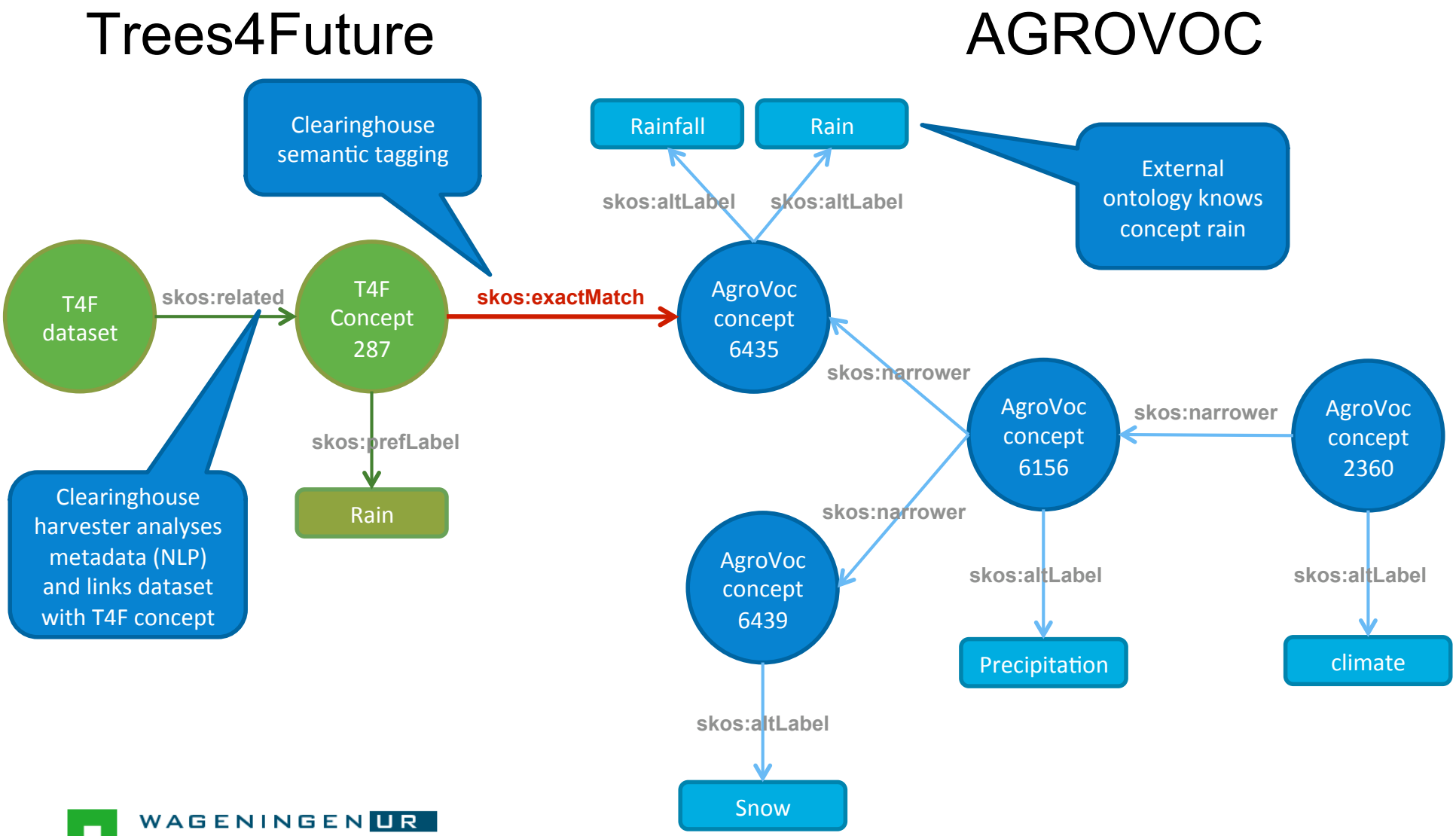
Results:15 total

N	Name (click to see metadata)	Description	Publisher	Score	Services	Source
1	Ef03 x Sl03	Crossing; Quercus robur	Austrian Institute of Technology - AIT; Stephan Gaubitzer; stephan.gaubitzer@ait.ac.at		Go to page	GO
2	L10Q	Association Population; Quercus petraea	INRA; Véronique Jorge; jorge@orleans.inra.fr		Go to page	GO
3	L12Q	Association Population; Quercus petraea	INRA; Véronique Jorge; jorge@orleans.inra.fr		Go to page	GO
4	L14Q	Association Population; Quercus petraea	INRA; Véronique Jorge; jorge@orleans.inra.fr		Go to page	GO
5	L16Q	Association Population; Quercus petraea	INRA; Véronique Jorge; jorge@orleans.inra.fr		Go to page	GO
6	L1Q	Association Population; Quercus petraea	INRA; Véronique Jorge; jorge@orleans.inra.fr		Go to page	GO


Find rainfall data
Also look for data on precipitation, rain...
Trees4Future Clearinghouse
User Interface
query datasets
Semantic Search Layer
semantic search
External ontologies
Metadata repository
NLP, ontology linking
Catalogue Harvester
Using ontologies: link precipitation to rainfall
Collects metadata of distributed datasets
Data repository 1
provides access
Data sets
describes
Metadata catalogue
harvests
Data repository 2
provides access
Data sets
describes
harvests
Metadata catalogue
Describes the datasets e.g. keyword = precipitation

Example – semantic tagging & search

Climate dataset - metadata contains term 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?

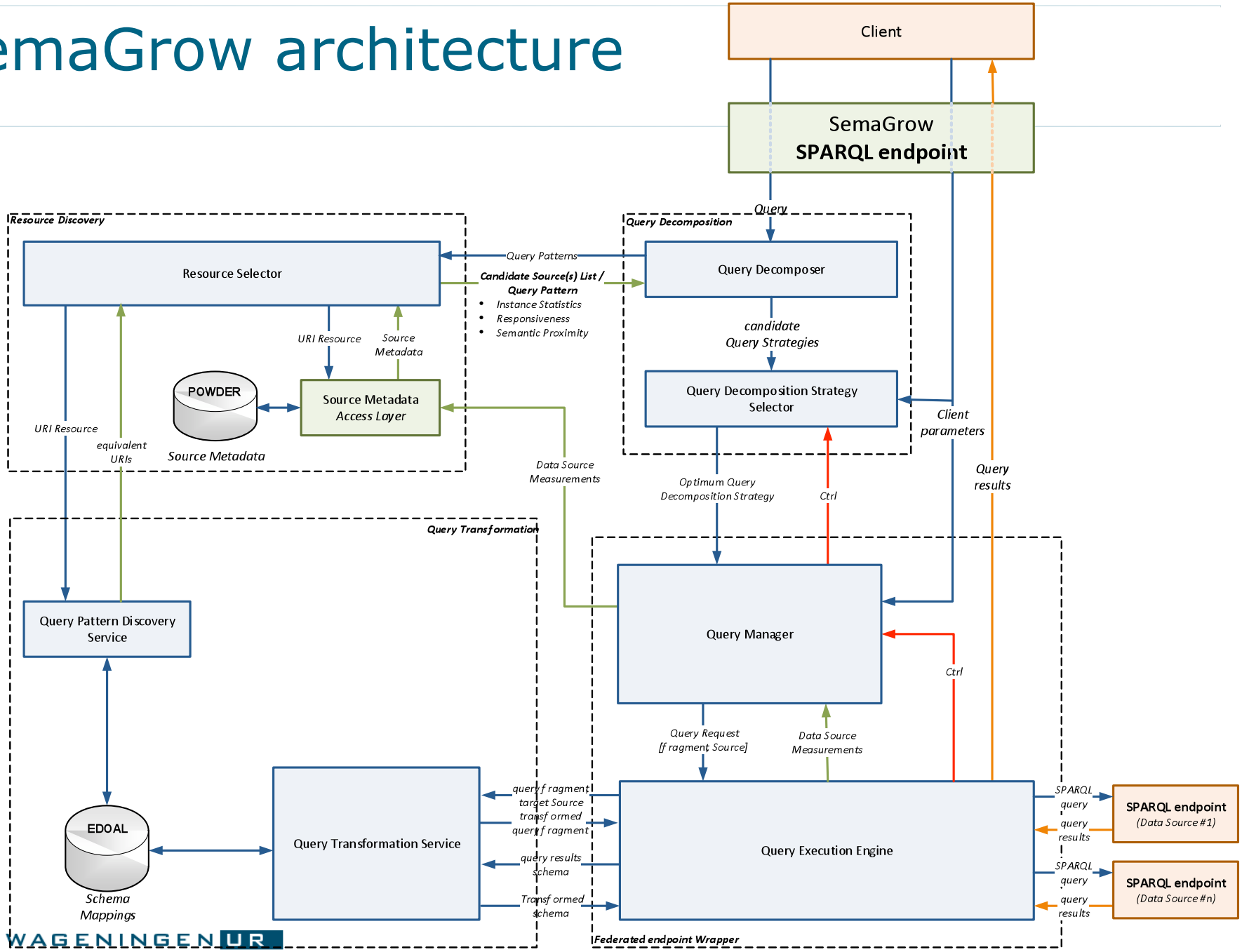
Example query:

- 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.

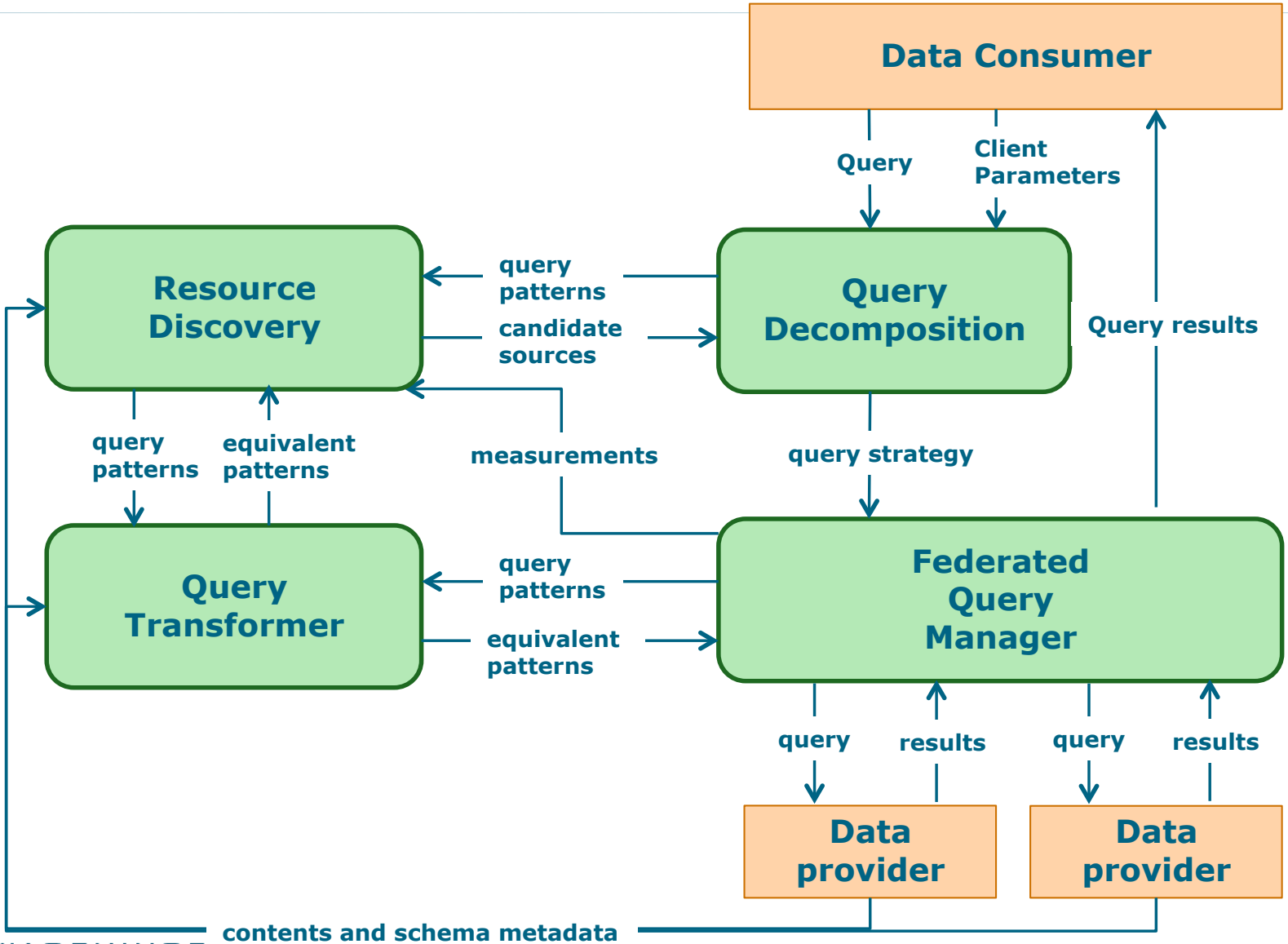
SemaGrow data sizes

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

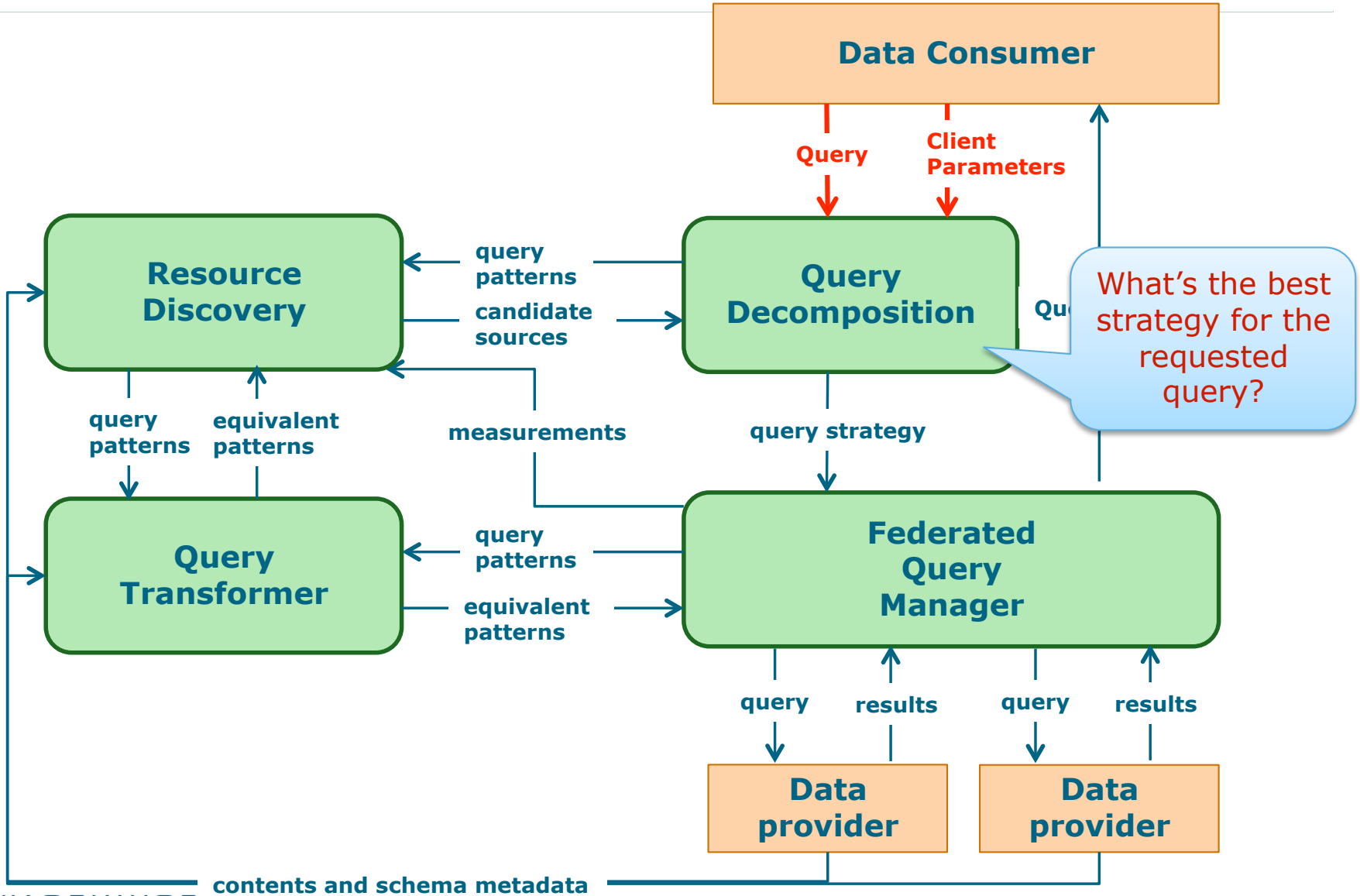
SemaGrow architecture



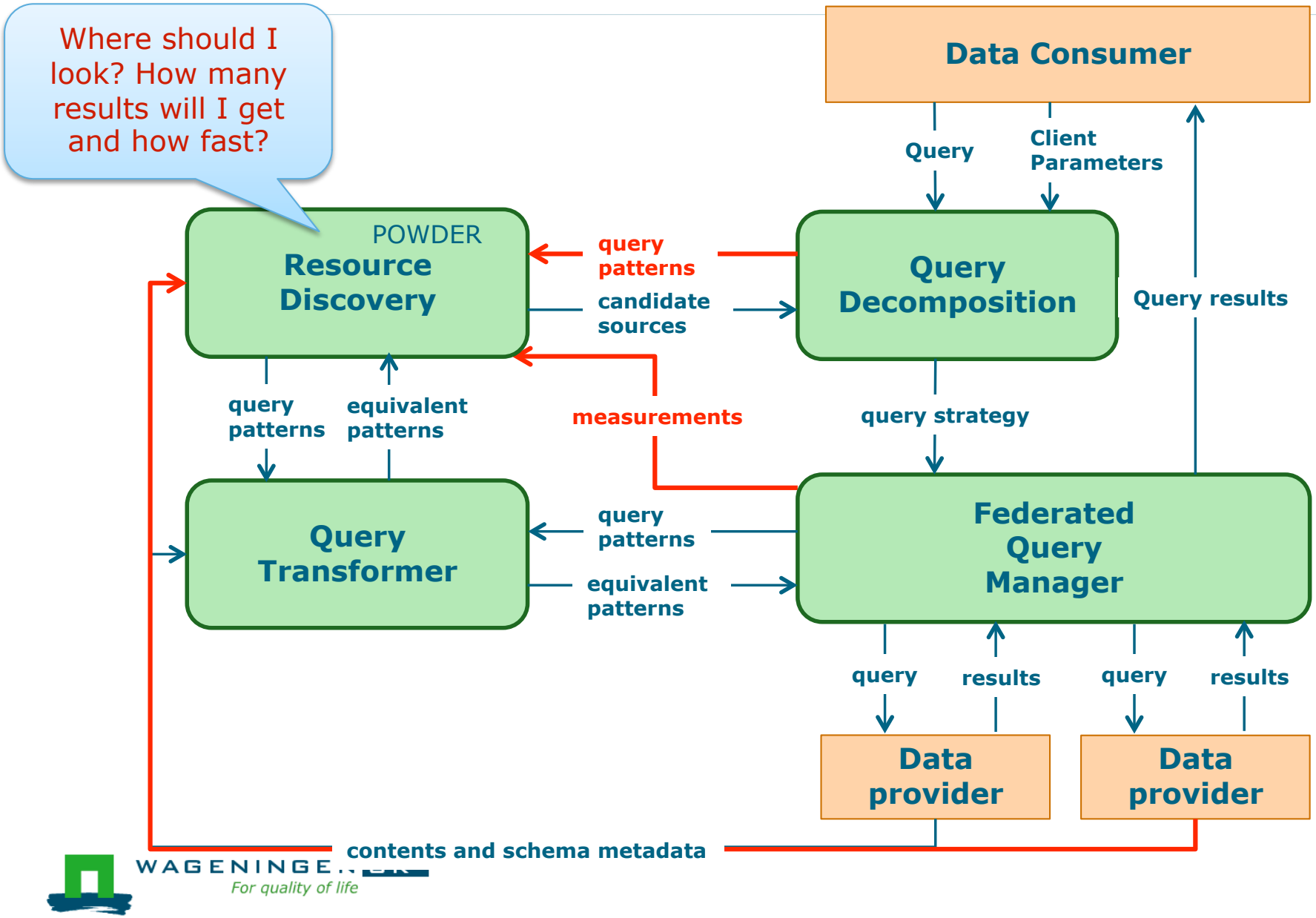
SemaGrow architecture



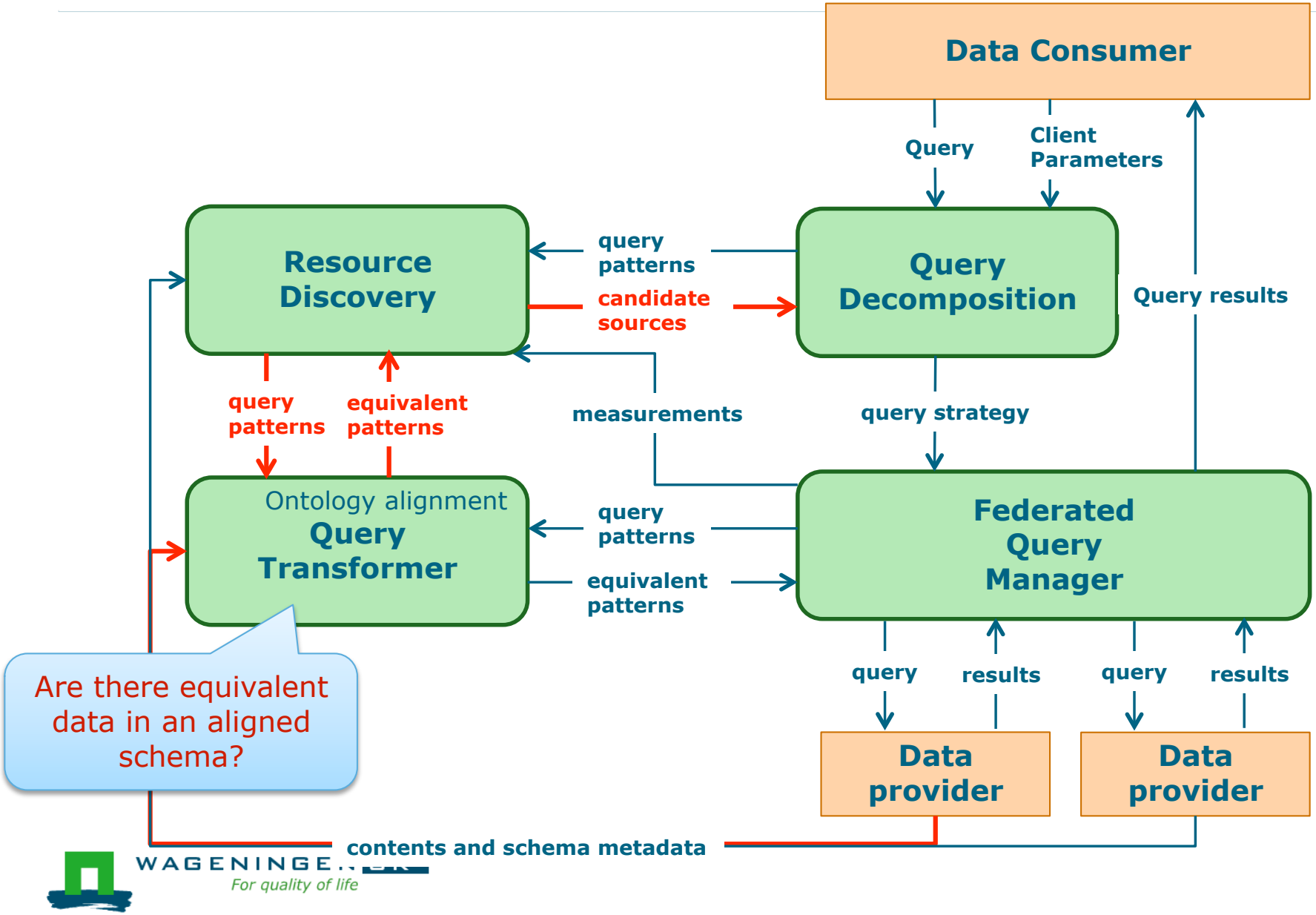
SemaGrow architecture



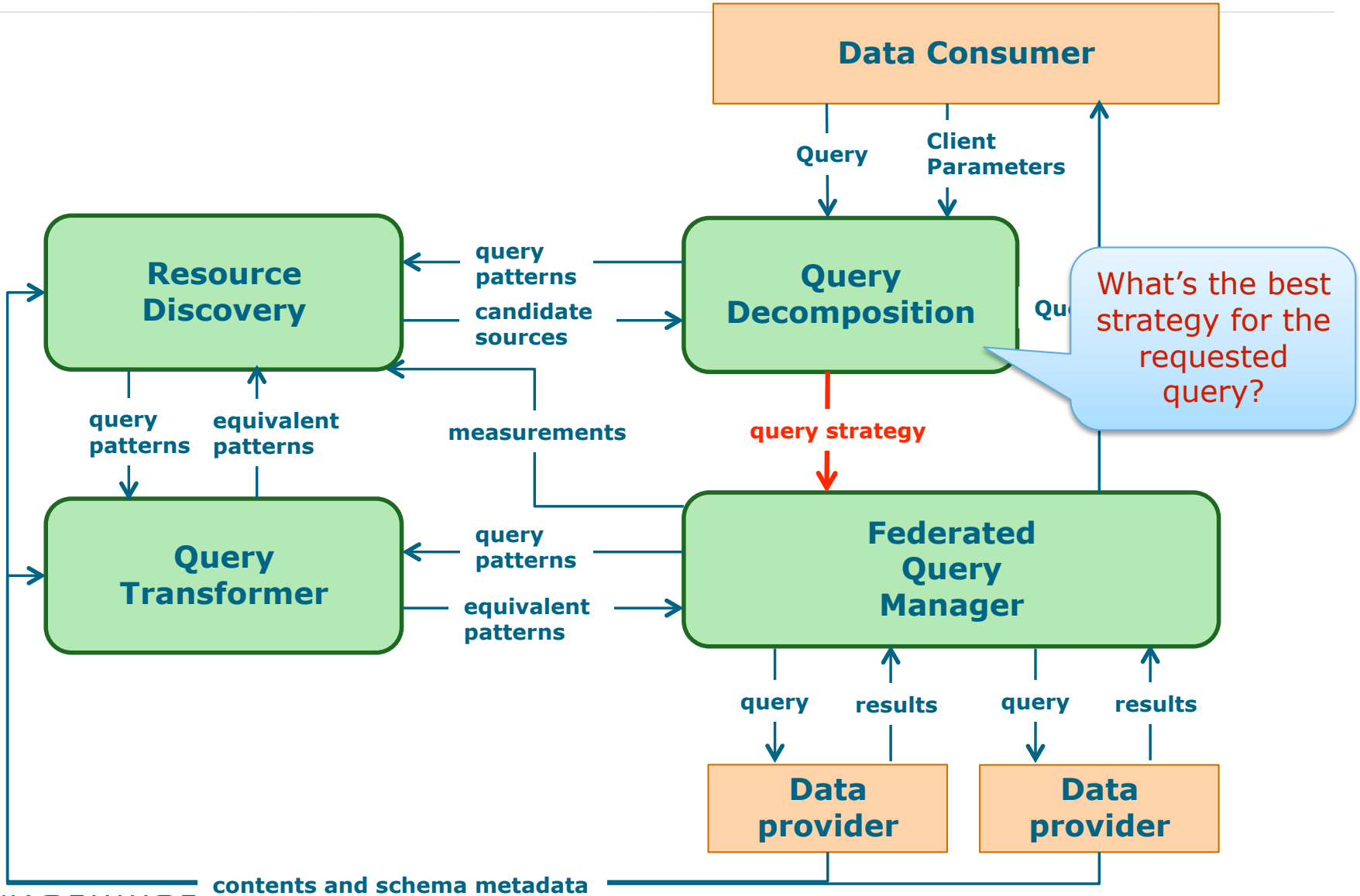
SemaGrow architecture



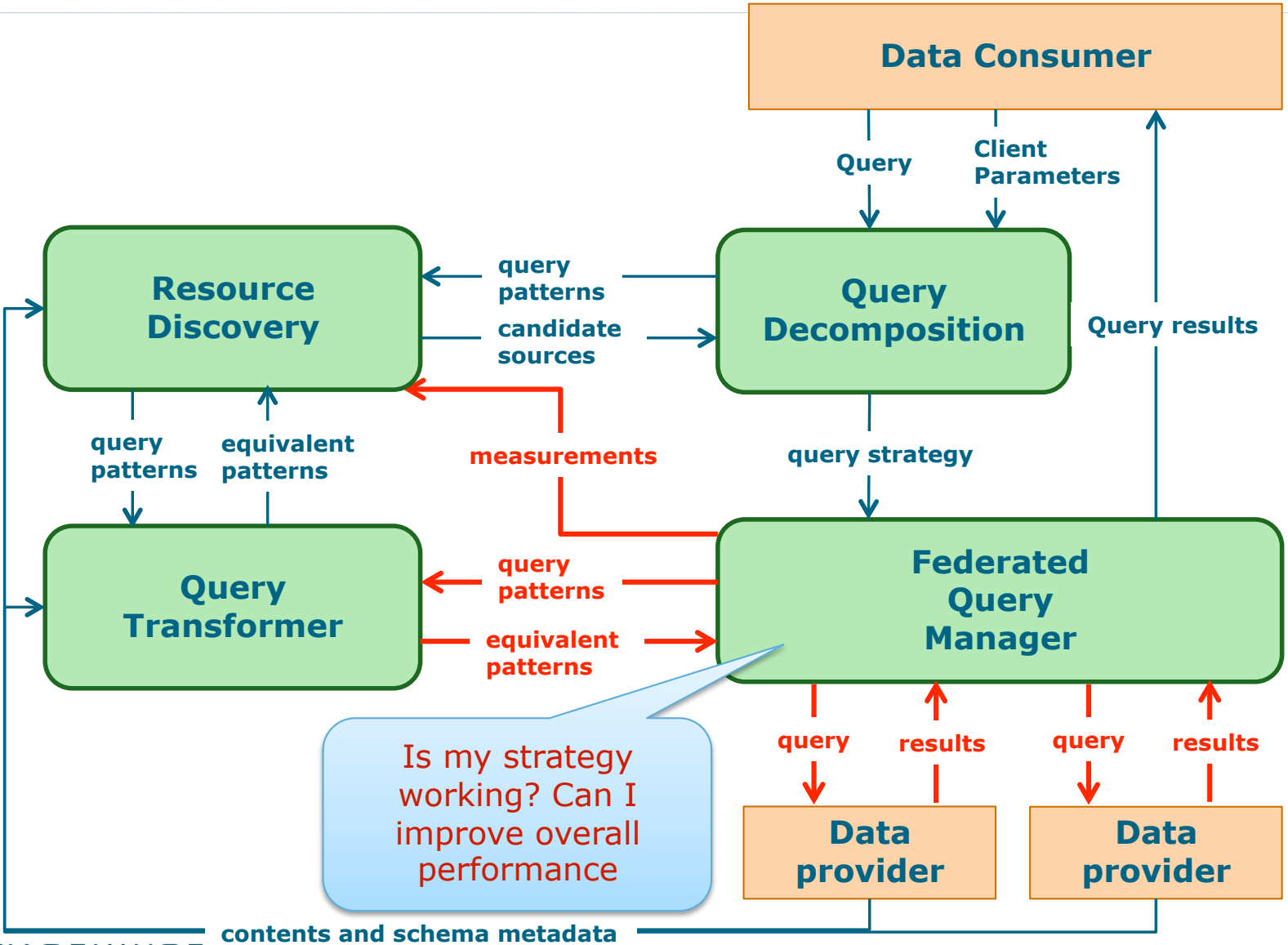
SemaGrow architecture



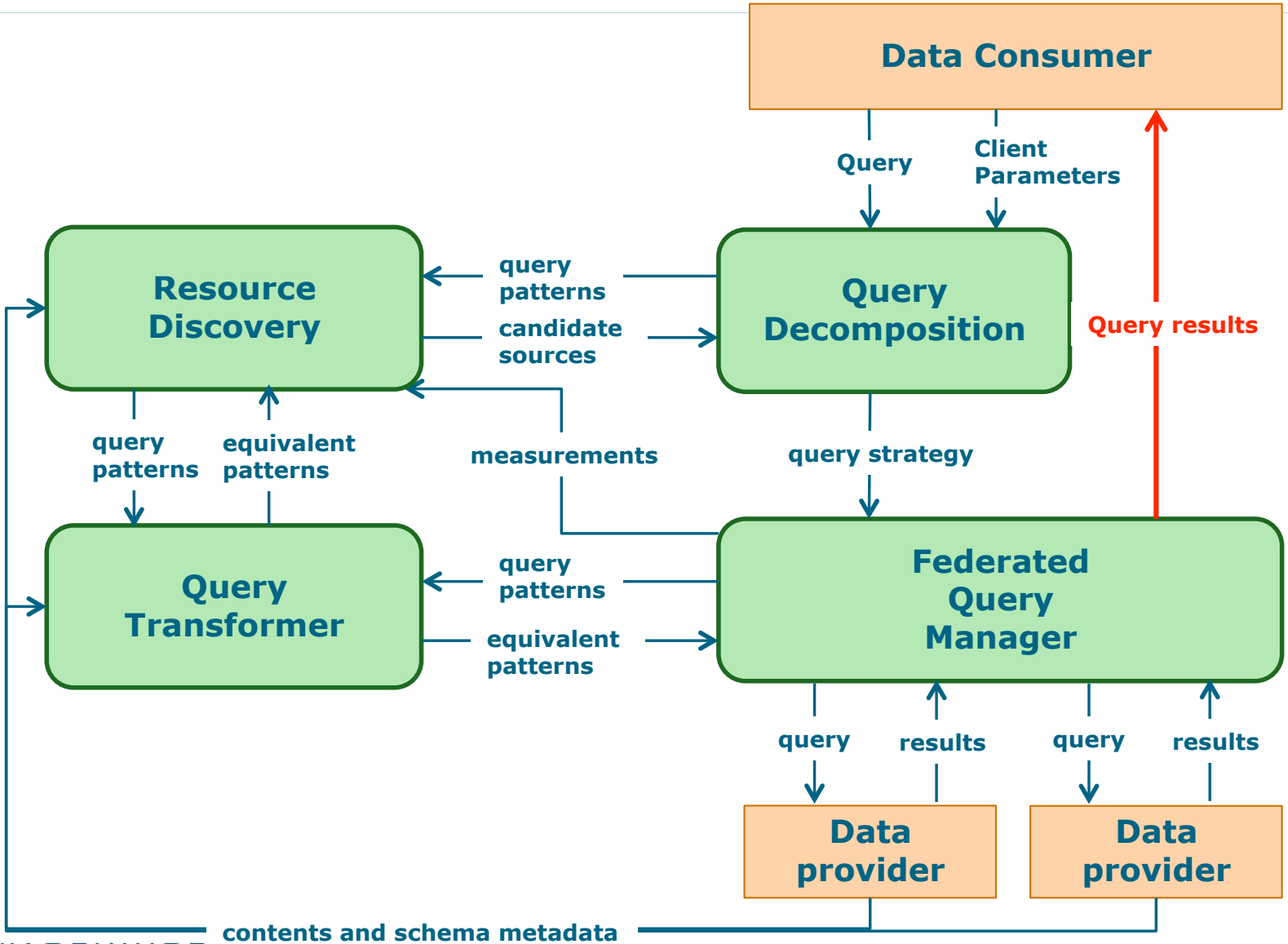
SemaGrow architecture



SemaGrow architecture



SemaGrow architecture



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!