Using Solr sharding for federated search of plant phenotypes?

EXCELERATE WP7

Dan Bolser
Ensembl Plants

plants.ensembl.org

slides:
http://tinyurl.com/solr2016

EMBL-EBI

Using ‘old-school’ Solr sharding to provide a distributed search for the transPLANT project

Dan Bolser, Ensembl Plants, EBI

http://transPLANTdb.eu

Overview

Won't talk about transPLANT

Won't talk about ELIXIR / EXCELERATE

Won't talk about SolrCloud

Will talk about 'old-school' Solr **distributed search**

Will talk about what we did with Solr in the transPLANT project

Will talk about EXCELERATE WP7

Why ‘old-school’?

Distributed Search:

- SolrCloud is a ‘more current’ approach to distributed search… introduced in Solr 4.0 and has many advancements that make distributed search easier
  - No time to cover SolrCloud properly here!
  - Main ‘problem’ with SolrCloud for us is ‘distributed indexing’.

Source: https://wiki.apache.org/solr/**DistributedSearch**

What is Distributed Search?

When an index becomes **too large** to fit on a single system, or when a single query takes **too long** to execute,
an index can be split into **multiple shards**, and
Solr can **query and merge** results across those shards.

Source: https://wiki.apache.org/solr/**DistributedSearch**

If single queries are currently fast enough and one simply wishes to expand the capacity (**queries/sec**) of the search system, then standard whole **index replication** should be used.

Solr can query and merge results across those shards

The current components that support distributed search are:

- Query component
  - Returns documents matching a query.
- Facet component
- Grouping component
  - From version Solr4.0.
  - Currently group.truncate and group.func are the only parameters that aren't supported for distributed searches.
- Highlighting component
- Stats component
- Spell Check Component
- Terms Component
- Term Vector Component
- Debug component

Distributed Searching Limitations

- Documents must have a unique key and the unique key must be stored (stored="true" in schema.xml)
  - The unique key field must be unique across all shards. If docs with duplicate unique keys are encountered, Solr will make an attempt to return valid results, but the behavior may be nondeterministic.
- No distributed IDF (inverse document frequency).
- Doesn't support Join.
- Doesn't support pivot facing.

Source: https://wiki.apache.org/solr/**DistributedSearch**

Distributed Indexing

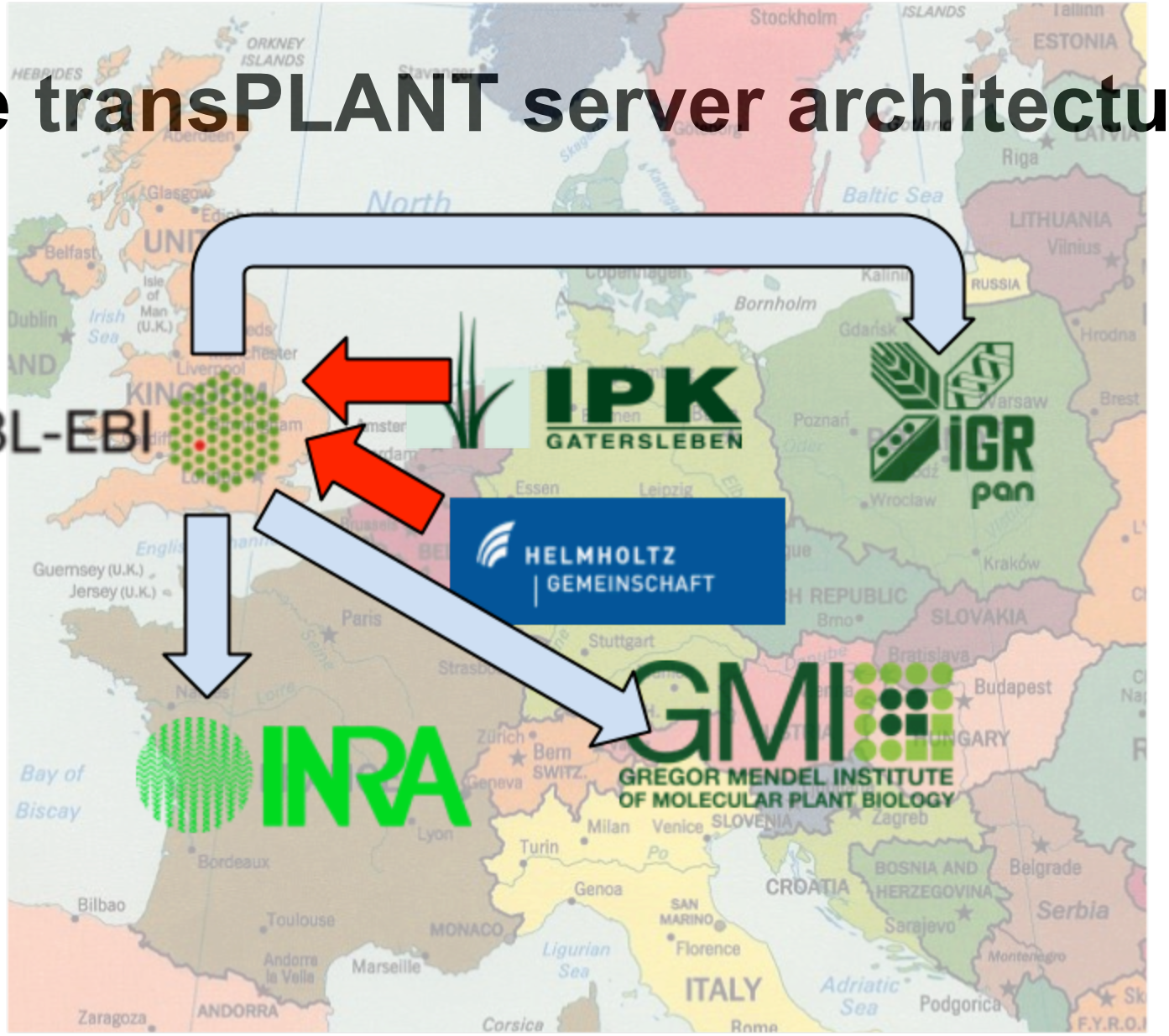
‘Old school’ sharding:

It's up to the user to distribute documents across shards.

SolrCloud implements distributed indexing and has various strategies available.

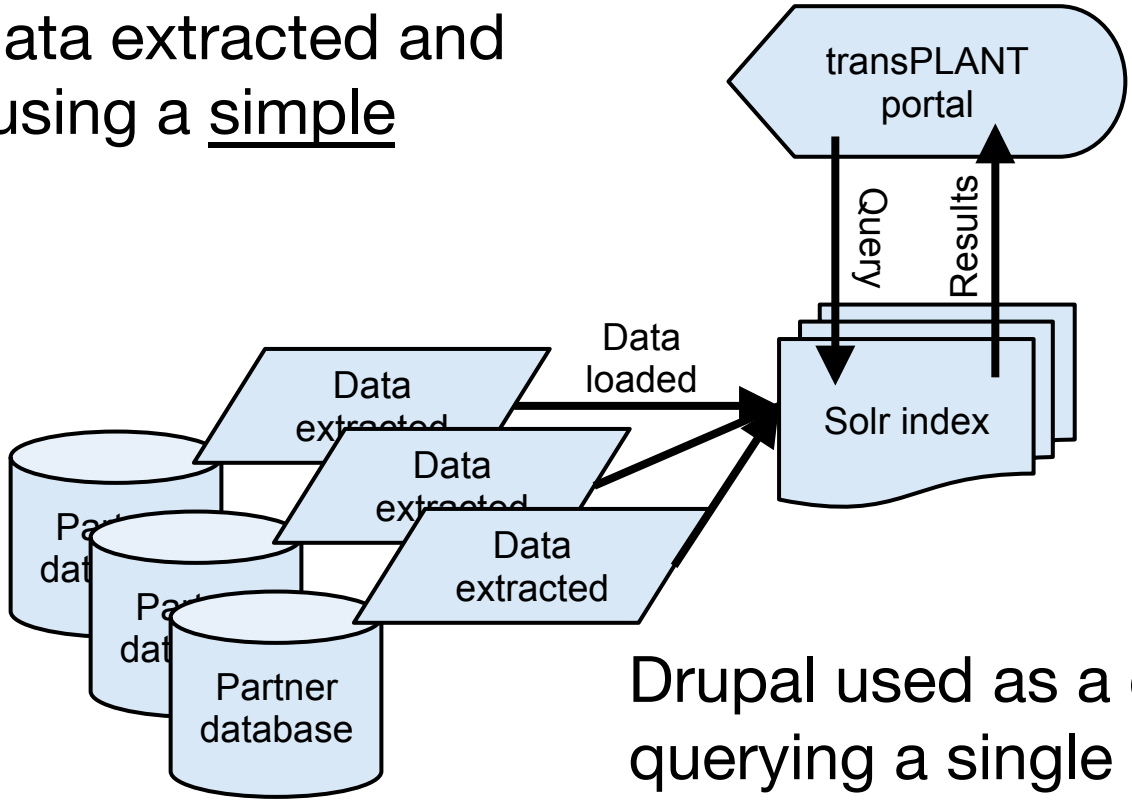
Source: https://wiki.apache.org/solr/**DistributedSearch**

The transPLANT server architecture



Our original search implementation

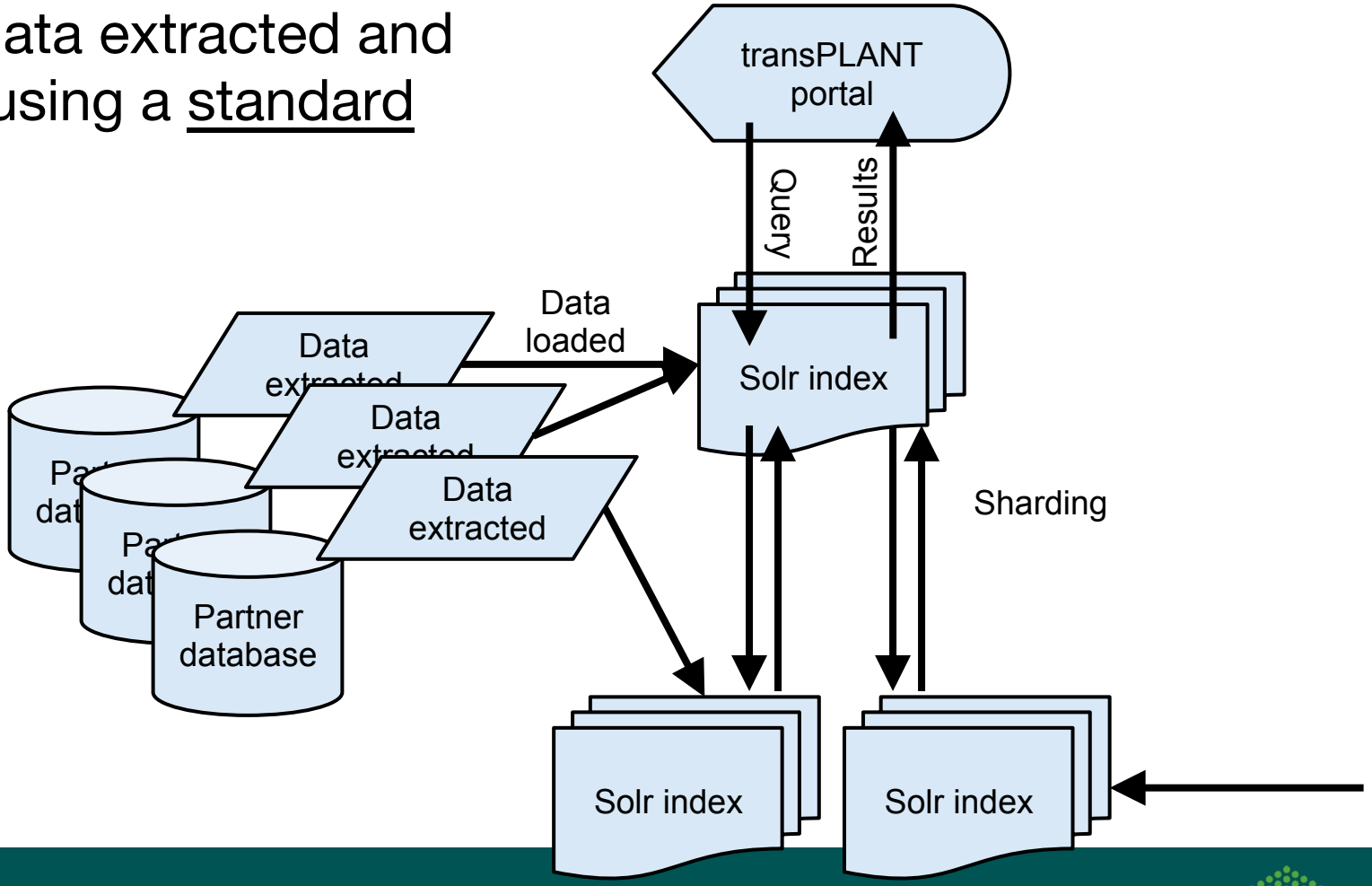
Partner data extracted and indexed using a simple schema



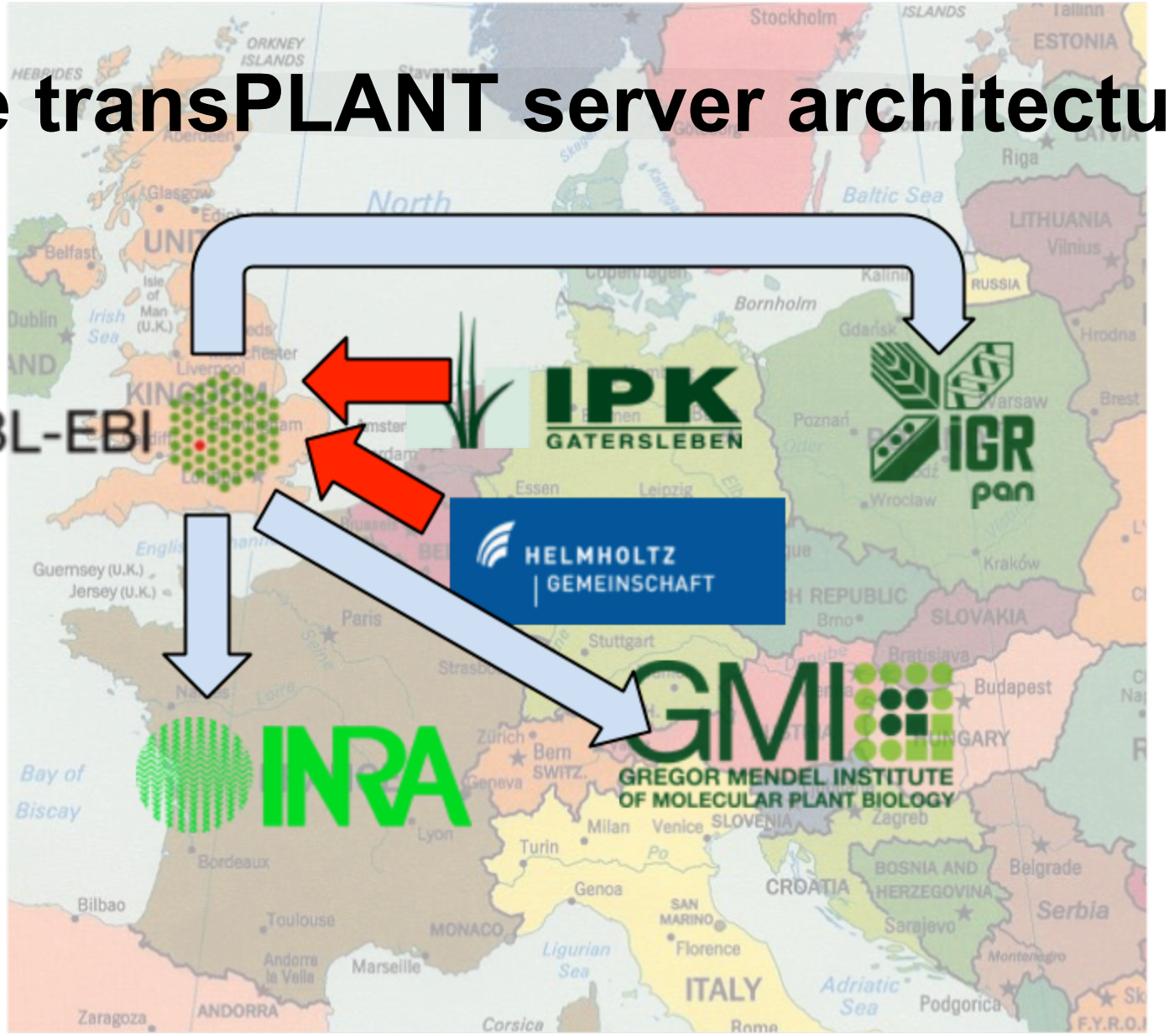
Drupal used as a client, querying a single Solr index

Truly distributed search

Partner data extracted and indexed using a standard schema

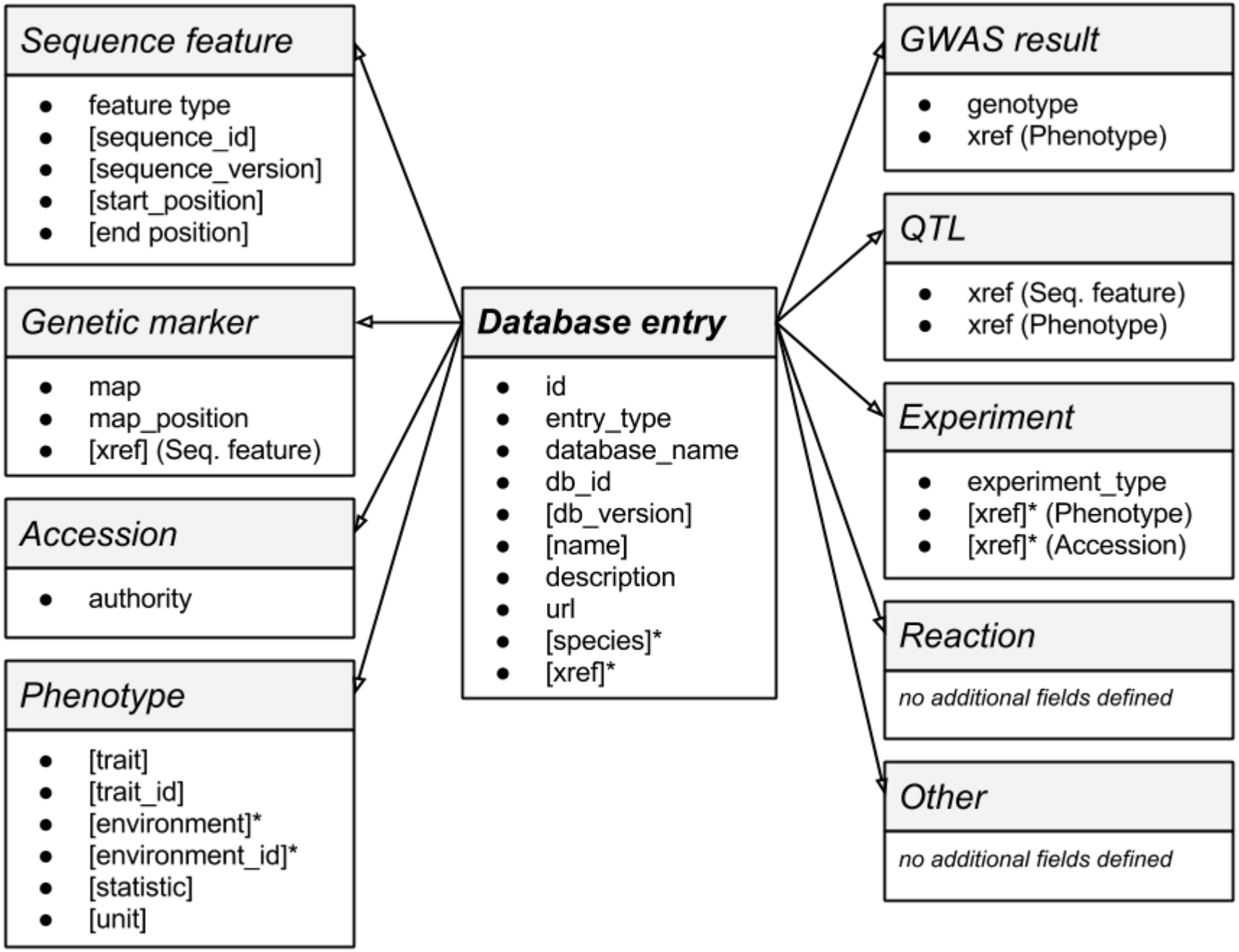


The transPLANT server architecture



Heterogeneous data…

Provider	Database name	Database description	Number of entries	Number of species	Entry types	Location of index
EBI	Ensembl Plants	Ensembl Plants is a genome-centric portal for plant species of scientific interest. It utilises reference genome sequences as a framework to integrate variant, functional, expression, marker and comparative data through a consistent set of interactive and programmatic interfaces.	1,523,622	38	Sequence feature (protein_coding, ncRNA)	UK
PAS	PolapgenDB		31	Hordeum vulgare	Phenotype, Accession, Experiment, Genetic marker, Sequence feature	Poland
MIPS	PlantsDB	PlantsDB is a database platform providing tools and views for the comparative analysis of plant genomes and transcriptome data.	263,401	6	Sequence feature (transcript)	UK
MIPS	CrowsNest	A comparative map viewer to visualize and investigate genome-wide chromsome organization as well as genome-wide synteny between two or more plant genomes.	140,138	4	Sequence feature	UK
IPK	OPTIMAS-DW	A comprehensive transcriptomics, metabolomics, ionomics, proteomics and phenomics data resource for maize.	33,919	Zea mays	Sequence feature	UK
IPK	MetaCrop	MetaCrop is a database that summarizes diverse information about metabolic pathways in crop plants and allows automatic export of information for the creation of detailed metabolic models.	586	>50	Reaction	UK
IPK	GBIS	GBIS/I allows to retrieve information from the German federal ex situ collection.	148,696	>50	Accession	UK
IPK	CR-EST	The Crop EST Database (CR-EST) is a public available online resource providing access to sequence, classification, clustering, and annotation data of crop EST projects at the IPK-Gatersleben.	218,927	6	Sequence feature (EST)	UK
INRA	GnpIS	GnpIS is a multispecies integrative information system	27,366	>50	Accession, Experiment	France
GMI	GWAPortal	GWAPortal is a resource for phenotypes and GWAS st	828	Arabidopsis thaliana	GWAS, Experement, Phenotype	Germany


Sequence feature
feature type
[sequence_id]
[sequence_version]
[start_position]
[end position]
Genetic marker
map
map_position
[xref] (Seq. feature)
Accession
authority
Phenotype
[trait]
[trait_id]
[environment]*
[environment_id]*
[statistic]
[unit]
Database entry
id
entry_type
database_name
db_id
[db_version]
[name]
description
url
[species]*
[xref]*
GWAS result
genotype
xref (Phenotype)
QTL
xref (Seq. feature)
xref (Phenotype)
Experiment
experiment_type
[xref]* (Phenotype)
[xref]* (Accession)
Reaction
no additional fields defined
Other
no additional fields defined

About the transPLANT schema

The schema is designed to be **simple** enough to accomodate results from **different and varied** sources, yet detailed enough to provide meaningful free text search and subsequent exploration of results via **faceting, filtering** and **rich-snippets**, should the client application chose to implement them.

The schema is proposed for the purpose of supporting integrated search over different data providers, to give users **a single point of entry into multiple databases**. As such, the schema isn't intended to be a detailed model of the underlying data, biology, or the scientific processes used to gather it.

```
  <requestHandler name="standard" class="solr.StandardRequestHandler" default="true" />

  <requestHandler name="pinkPony" class="solr.SearchHandler" default="true" startup="lazy">

    <lst name="defaults">

      <!-- Add our shard servers... -->
      <str name="shards">localhost:${jetty.port:8983}/solr/transPlant-EBI,localhost:${jetty.port:8983}/solr/transPlant-IPK,localhost:${jetty.port:8983}/solr/transPlant-MIPS,urgi.versailles.inra.fr/solr/transplant-0.2-pre,cropnet.pl/solr-transplant-0.2,gwas.gmi.oeaw.ac.at:8983/solr</str>

      <!-- Shards moved from local cores to remote cores:

           PAS:
           cropnet.pl/solr-transplant-0.2
           localhost:${jetty.port:8983}/solr/transPlant-PAS

           URGI:
           urgi.versailles.inra.fr/solr/transplant-0.2-pre
           localhost:${jetty.port:8983}/solr/transPlant-URGI

           GMI:
           gwas.gmi.oeaw.ac.at:8983/solr

       -->

      <!-- Print a summary of results per shard -->
      <str name="shards.info">true</str>

      <!-- Continue if one or more shards are down -->
      <str name="shards.tolerant">true</str>

    </lst>

  </requestHandler>

  <requestHandler name="/admin/" class="org.apache.solr.handler.admin.AdminHandlers" />
  <admin>
    <defaultQuery>*:*</defaultQuery>
  </admin>
```

Some example queries...

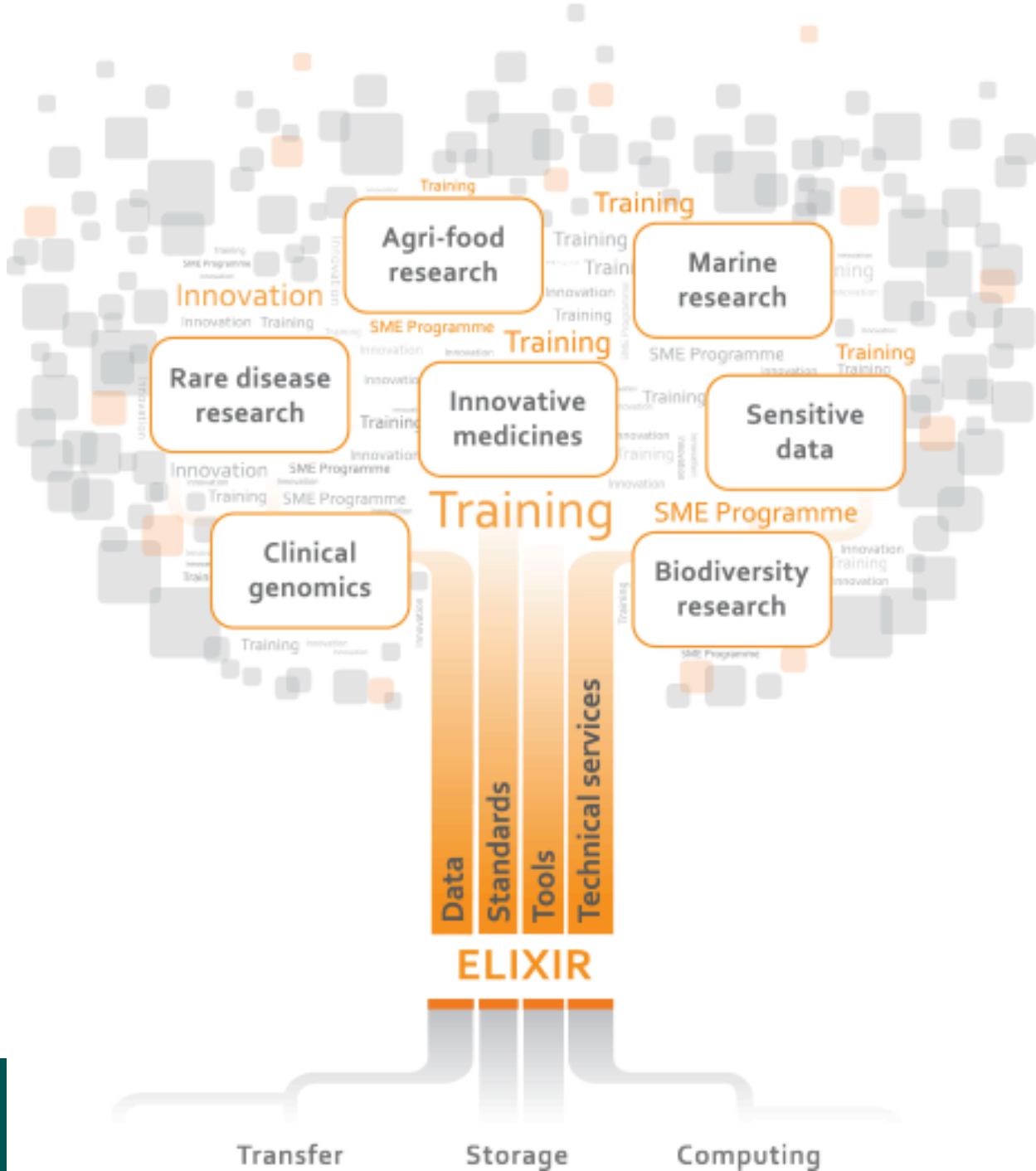
http://solr.transplantdb.eu/solr/transPlant-EBI/select?
q=rubisco&
shards=
    solr.transplantdb.eu/solr/transPlant-EBI,
    solr.transplantdb.eu/solr/transPlant-IPK,
    solr.transplantdb.eu/solr/transPlant-MIPS,
    urgi.versailles.inra.fr/solr/URGI,
    cropnet.pl/solr-transplant-0.2,
    gwas.gmi.oeaw.ac.at:8983/solr&
shards.info=true&
wt=json&indent=true

http://www.transplantdb.eu/search/transPLANT/rubisco

schema.xml

https://github.com/
dbolser-ebi/
transPLANT-Search/
blob/version-0.2/
solr/transPlant-EBI/conf/
schema.xml

ELIXIR


Agri-food research
Marine research
Rare disease research
Innovative medicines
Sensitive data
Clinical genomics
Biodiversity research
Training
SME Programme
Innovation
Data
Standards
Tools
Technical services
ELIXIR
Transfer
Storage
Computing

ELIXIR Infrastructure = Nodes + ELIXIR Platforms

Data
Sustain core data resources

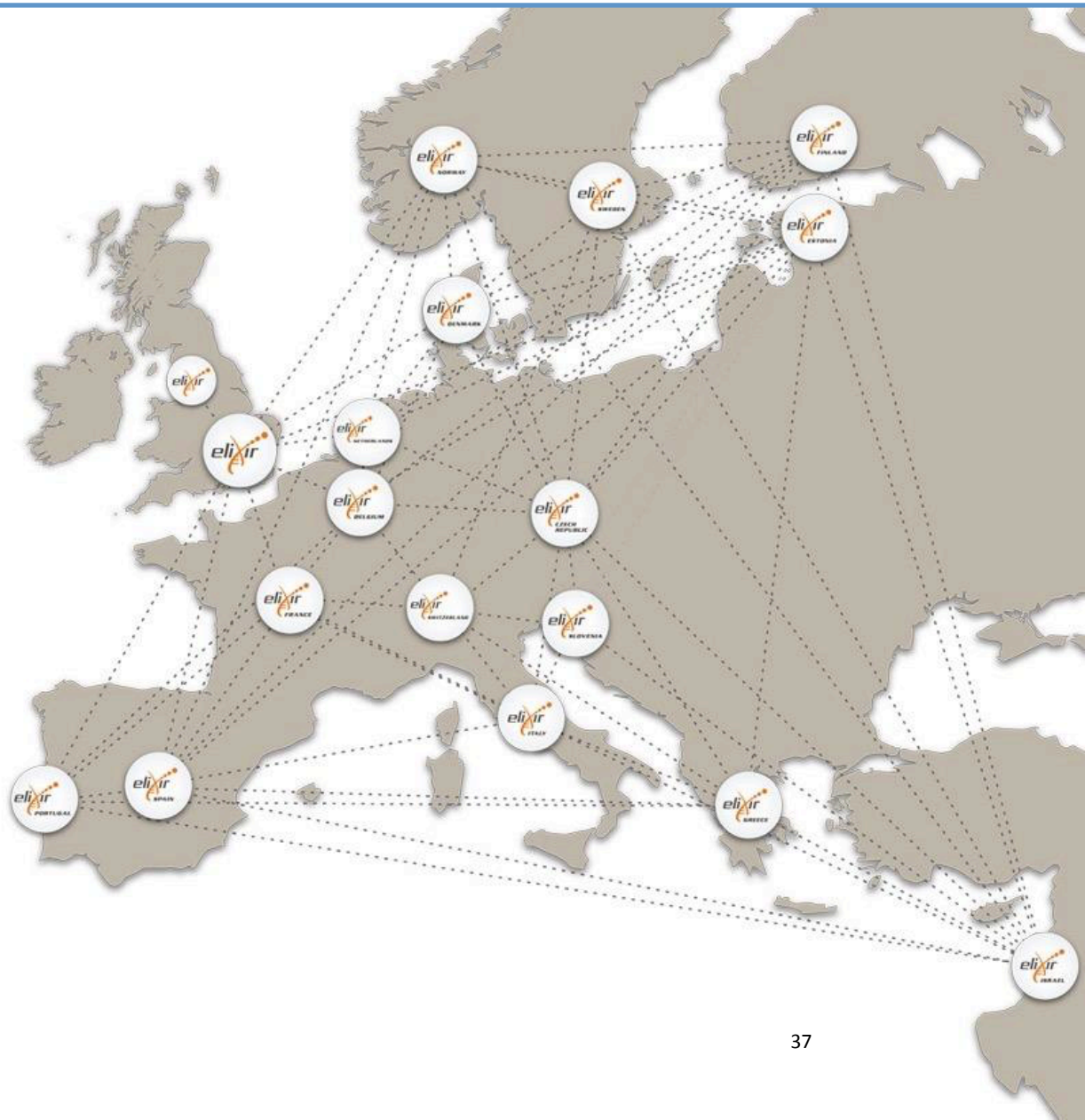
Tools
Services & connectors to drive access and exploitation

Interoperability
Integration and interoperability of data and services

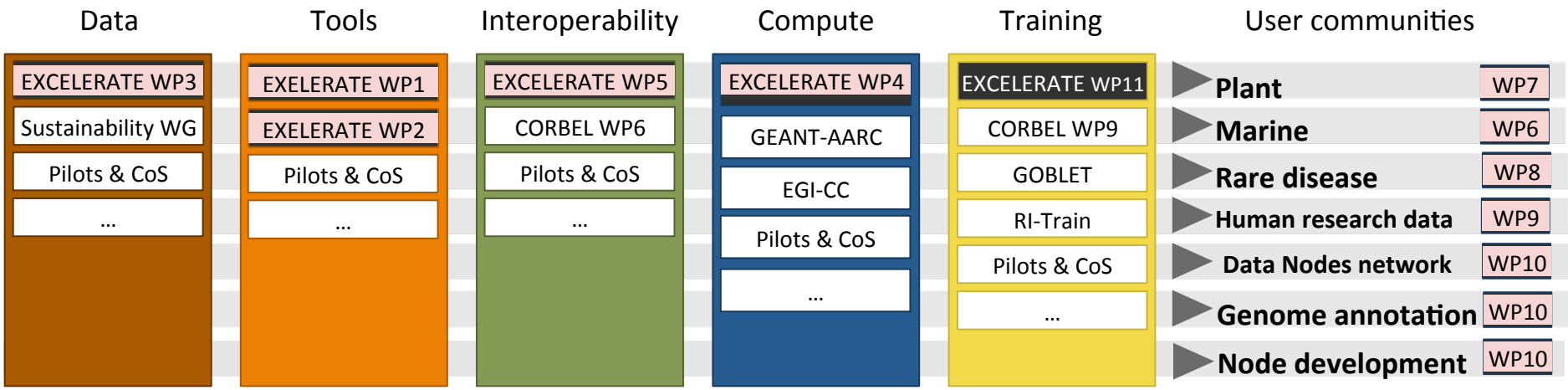
Compute
Access, Exchange & Compute (incl. sensitive data)

Training
Professional skills for managing and exploiting data

User Communities
Infrastructure usage and impact



User-driven implementation of ELIXIR through ELIXIR-EXCELERATE



ELIXIR Operational Management: Leadership team for each Platform (+ User community WP)

- Who?
  - Each platform lead by **EXCELERATE work package co-chairs**
  - Includes additional key leaders as required (e.g. CORBEL WPL)

- Role and remit?
  - **Lead overall implementation of ELIXIR Platform**
  - Empowered, resourced and accountable for platform delivery towards agreed strategy and objectives
  - Lead EXCELERATE work package
  - Coordinate efforts across grants to ensure synergies within ELIXIR
  - Propose activities for annual Workplan to be funded by ELIXIR core technical budget (Pilot actions and Commissioned Services)

- Accountable to?
  - **ELIXIR HoN**
  - September meeting (annual): Peer-review, prioritise and agree the annual ELIXIR Workplan including Pilots & Commissioned Services
  - Workplan submitted to ELIXIR Board by 1 October (annually) together with Budget proposal

Tasks

- Task 1 Development/adoption of appropriate controlled vocabularies for annotating plant phenotypic data
- Task 2 Annotation of key plant phenotypic data sets with agreed controlled vocabularies
- Task 3 Submission of exemplar genomic and phenotypic data sets to appropriate public repositories
- Task 4 Development and implementation of agreed public APIs for access to data in participating repositories and exposure via public computational infrastructures

Minimal objectives

- A common API to distributed data
  - We do not have a committed funding stream to maintain a unified European Plant phenotype database
  - But participating institutions are storing, organising and publishing plant phenotypic data – can we provide a single point of entry?
    - Emphasis on API rather than building a portal – but API will allow individual partners to build their own interactive interfaces over the services

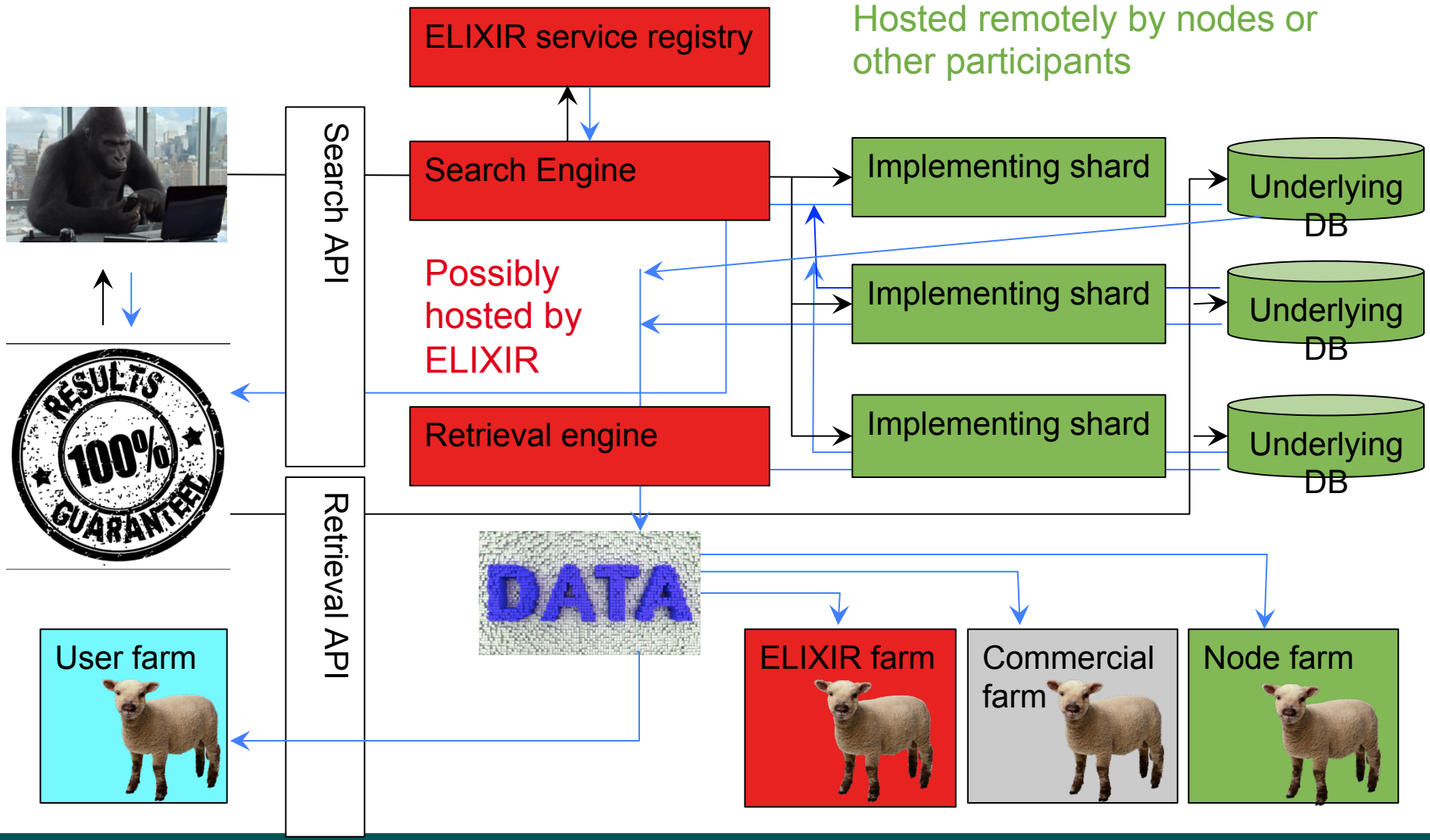
Minimal objectives

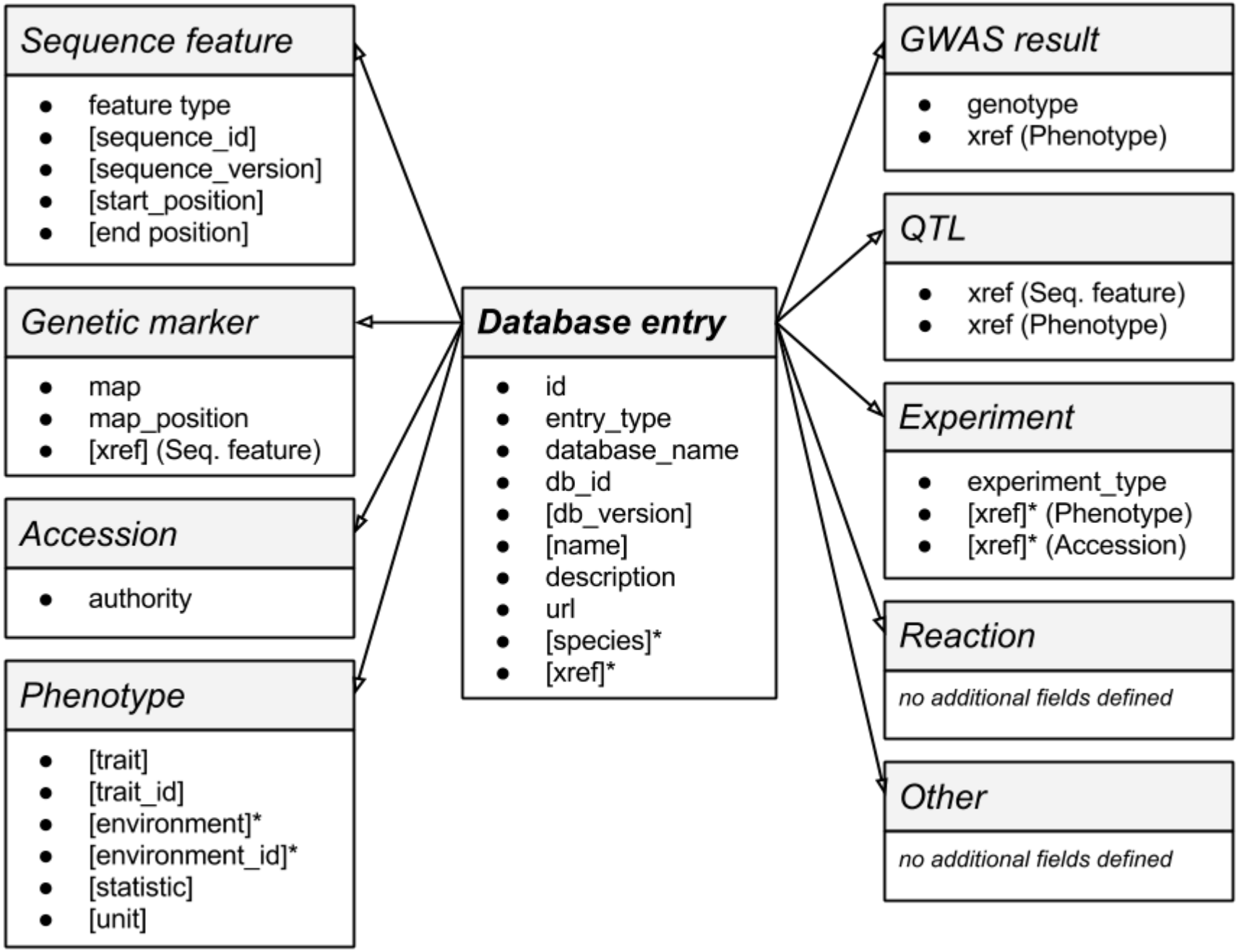
- Data discovery not data mining
  - Phenotypic data is very complex and diverse and it's probably not reasonable to expect us to prepare our data in such a way that people can automatically perform computational analysis across multiple data sets without studying/preparing the data
  - But we should make it possible for people to discover what data sets they might want to work with, even when these are stored in a distributed fashion

Data management

- Distributed model
  - Data is varied – no budget to build a centralised archive, instead looking at building lightweight services integrating over independently established archives
  - Data transfer between partners is likely to be small – simply the exposure of relevant meta-data for search service, possibly on request

Workflow (user data access)




Sequence feature
feature type
[sequence_id]
[sequence_version]
[start_position]
[end position]
Genetic marker
map
map_position
[xref] (Seq. feature)
Accession
authority
Phenotype
[trait]
[trait_id]
[environment]*
[environment_id]*
[statistic]
[unit]
Database entry
id
entry_type
database_name
db_id
[db_version]
[name]
description
url
[species]*
[xref]*
GWAS result
genotype
xref (Phenotype)
QTL
xref (Seq. feature)
xref (Phenotype)
Experiment
experiment_type
[xref]* (Phenotype)
[xref]* (Accession)
Reaction
no additional fields defined
Other
no additional fields defined

Thanks

http://tinyurl.com/solr2016

dbolser@ebi.ac.uk