EURISCO

Present and future

Stephan Weise
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK),
Gatersleben, Germany

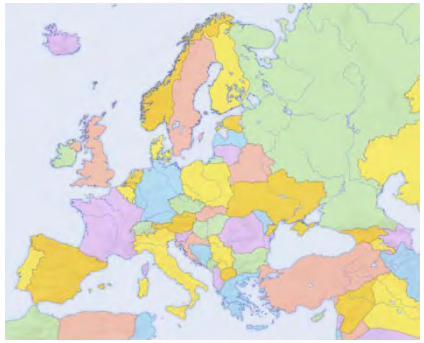
Workshop *"Tailoring the Documentation of Plant Genetic Resources in Europe to the Needs of the User",* Prague, Czech Republic, 20–22 May 2014

Outline

- Current status
- EURISCO transfer and reengineering
- Future developments

Introduction

- Development of European information system for plant genetic resources
  - Started in 2001 (EU project EPGRIS)
  - EURISCO: Search catalogue for *ex situ* collections; available since 2003
  - National collections represented by National Inventories (NIs)
  - 39 NIs involved (Nordic Countries → NGB)
  - Network of NFPs links NIs ↔ EURISCO

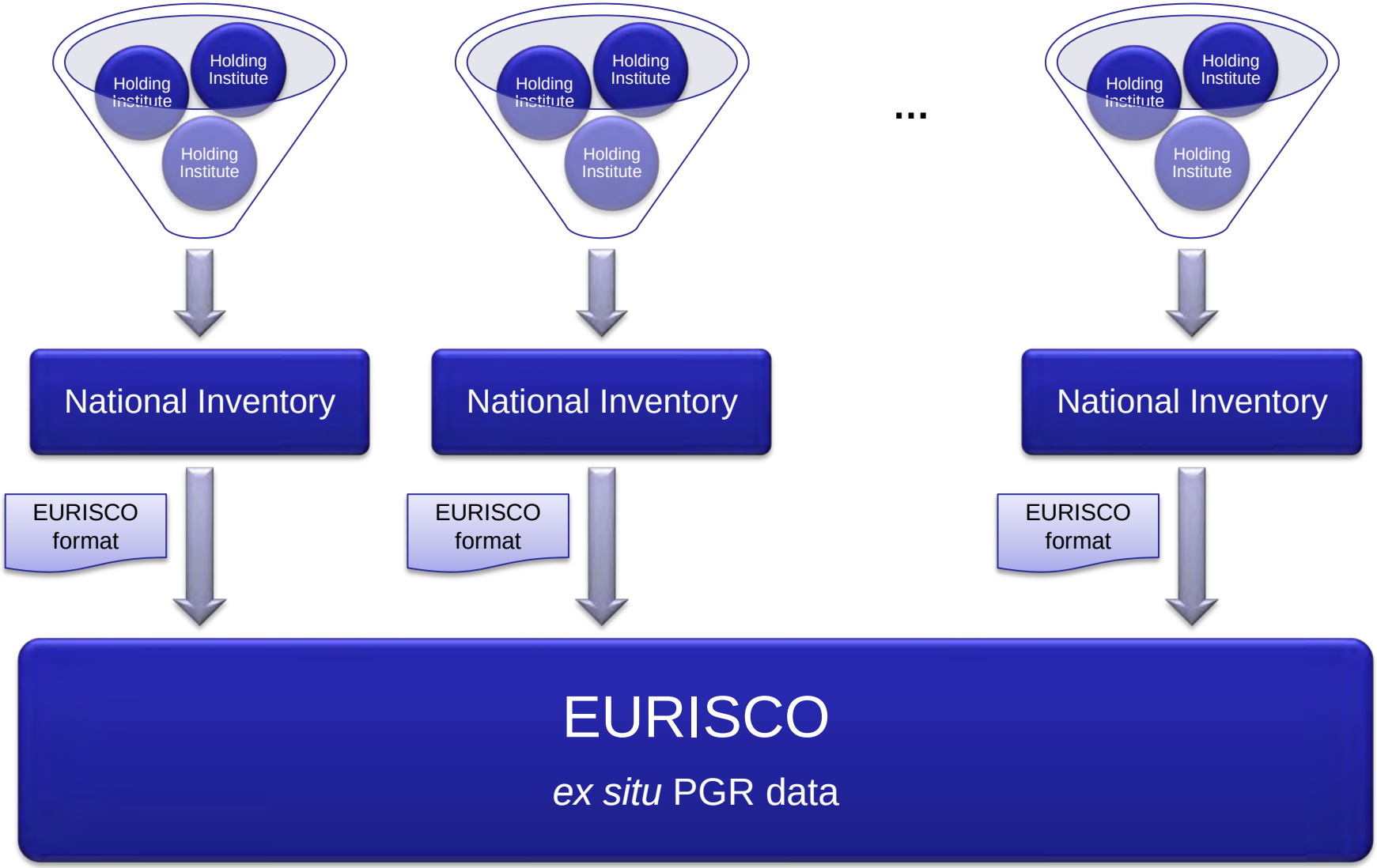


https://upload.wikimedia.org/wikipedia/commons/8/81/Europe_countries_map_2.png

Current status

- 39 National Inventories
- 1,074,965 accessions
- 5,885 genera
- 39,088 species (different combinations genus + species)
- ~9,000 site visits per year (as of 2012)

Current status



Outline

- Current status
- EURISCO transfer and reengineering
- Future developments

EURISCO transfer

- October 2012:
  - Request for tenders for hosting EURISCO
- March 2013:
  - IPK won the bid
- May and October 2013
  - Preparatory meetings in Gatersleben and Rome
- Since 15th April 2014:
  - Sub-contract with Bioversity International
  - Transfer of the EURISCO responsibilities started

Preparatory meetings

- Challenges:
  - Outdated systems
  - Insufficient technical documentation
  - High costs for transfer of as-is status

- Decisions:
  - →No transfer of existing solutions
  - →New development from scratch

Reengineering I

- EURISCO currently maintained in parallel at IPK and Bioversity

- Already done:
  - Analysis of current web application
  - Reengineering of database schema for web application
  - Import (and cleansing) of current data set
    - MySQL → Oracle RDBMS

Powered by MySQL®
http://www.mysql.com

ORACLE®
http://www.oracle.com

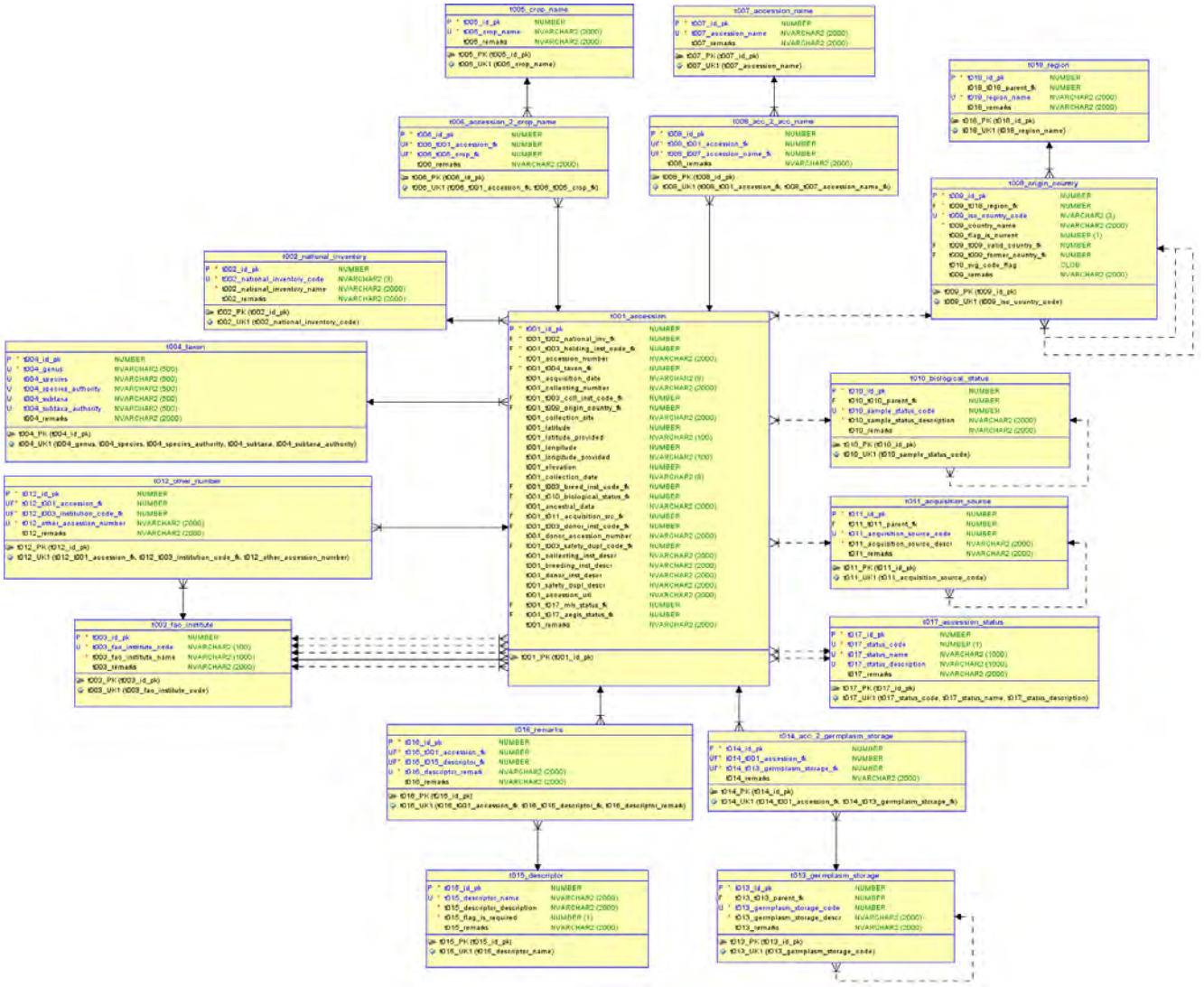
Reengineering II

- Under implementation:
  - New web application for searching EURISCO data
  - PL/SQL for backend; APEX for rendering
  - Prototype with basic functionality existing
  - Will be improved continuously
    - Content
    - Additional features
    - Performance tuning etc.

Reengineering III

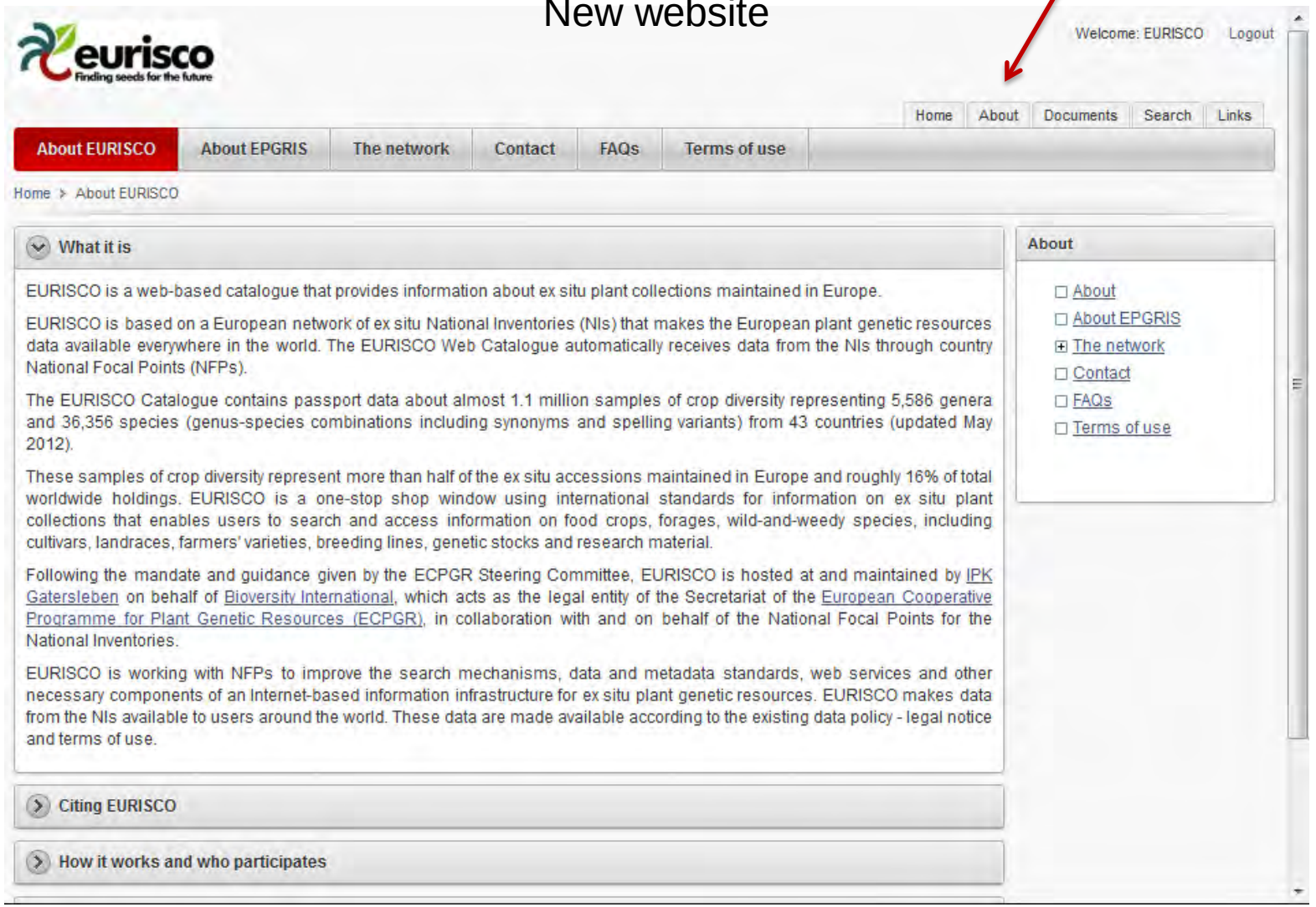
New database schema

- 18 tables + sequences
- 28 relationships
- 202 attributes
- 36 triggers
- 120 indexes
- For web application
  - Additional tables
  - Materialised views
  - PL/SQL procedures
  - JavaScript functions
  - …

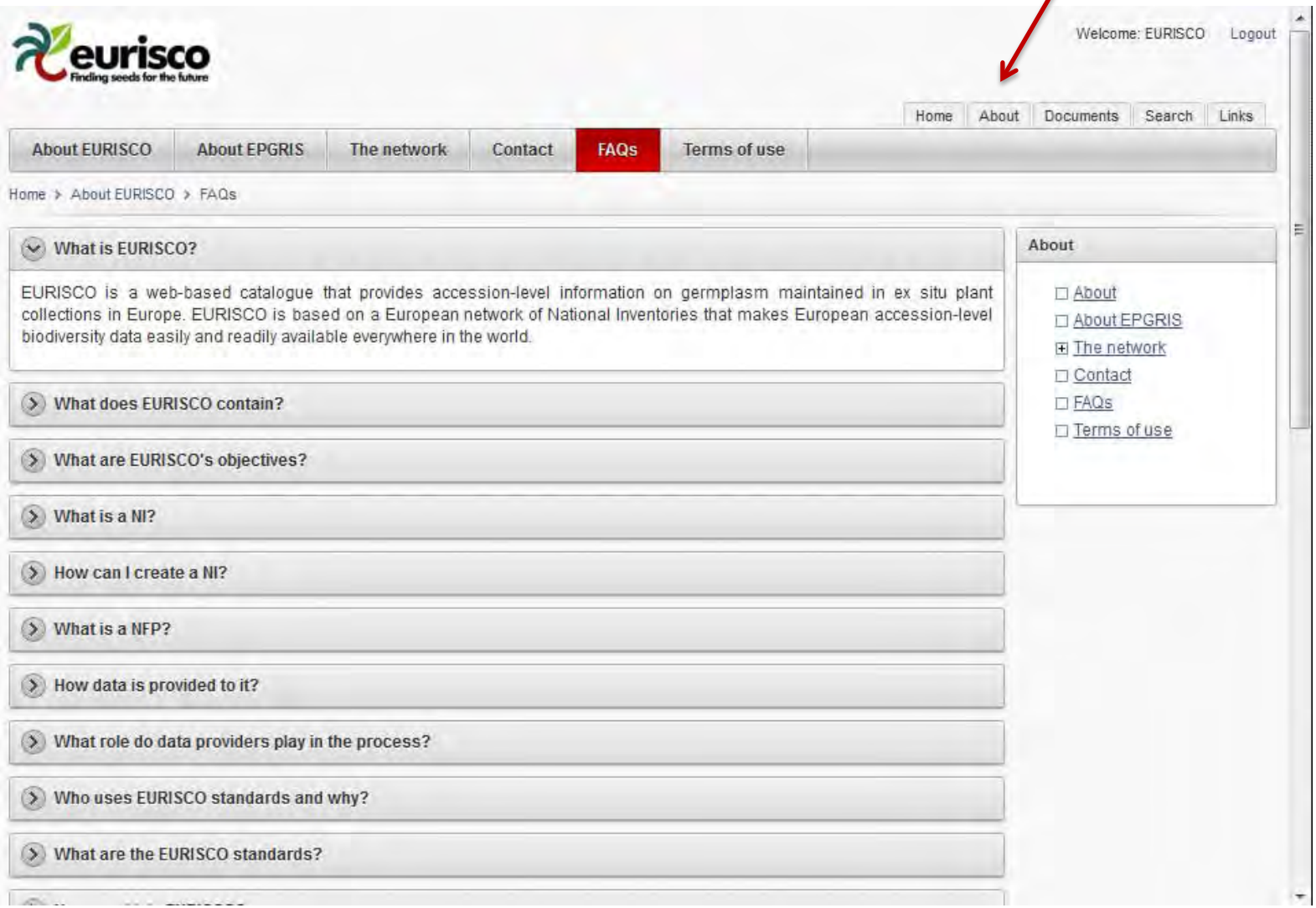


Reengineering IV

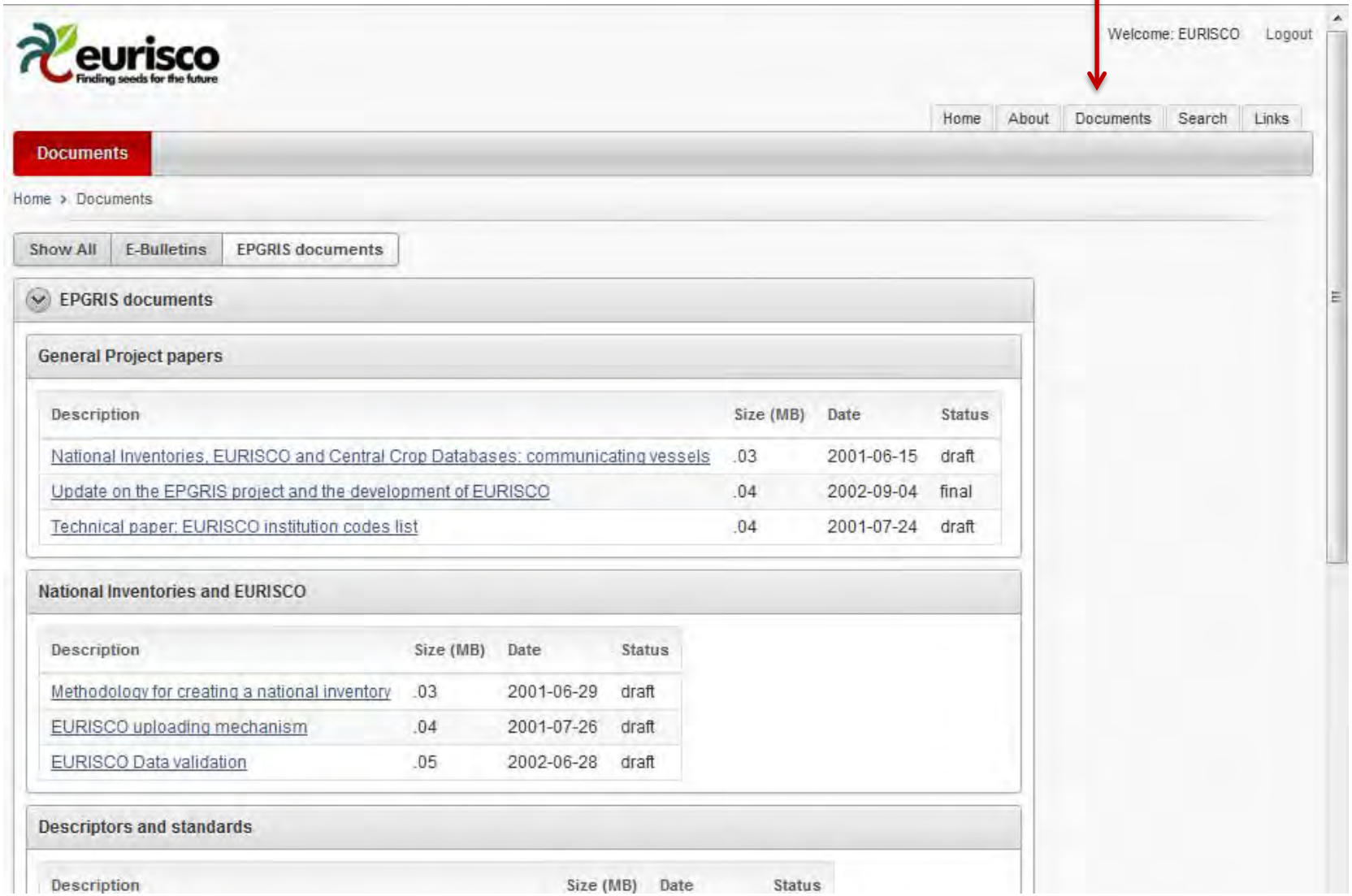
New website



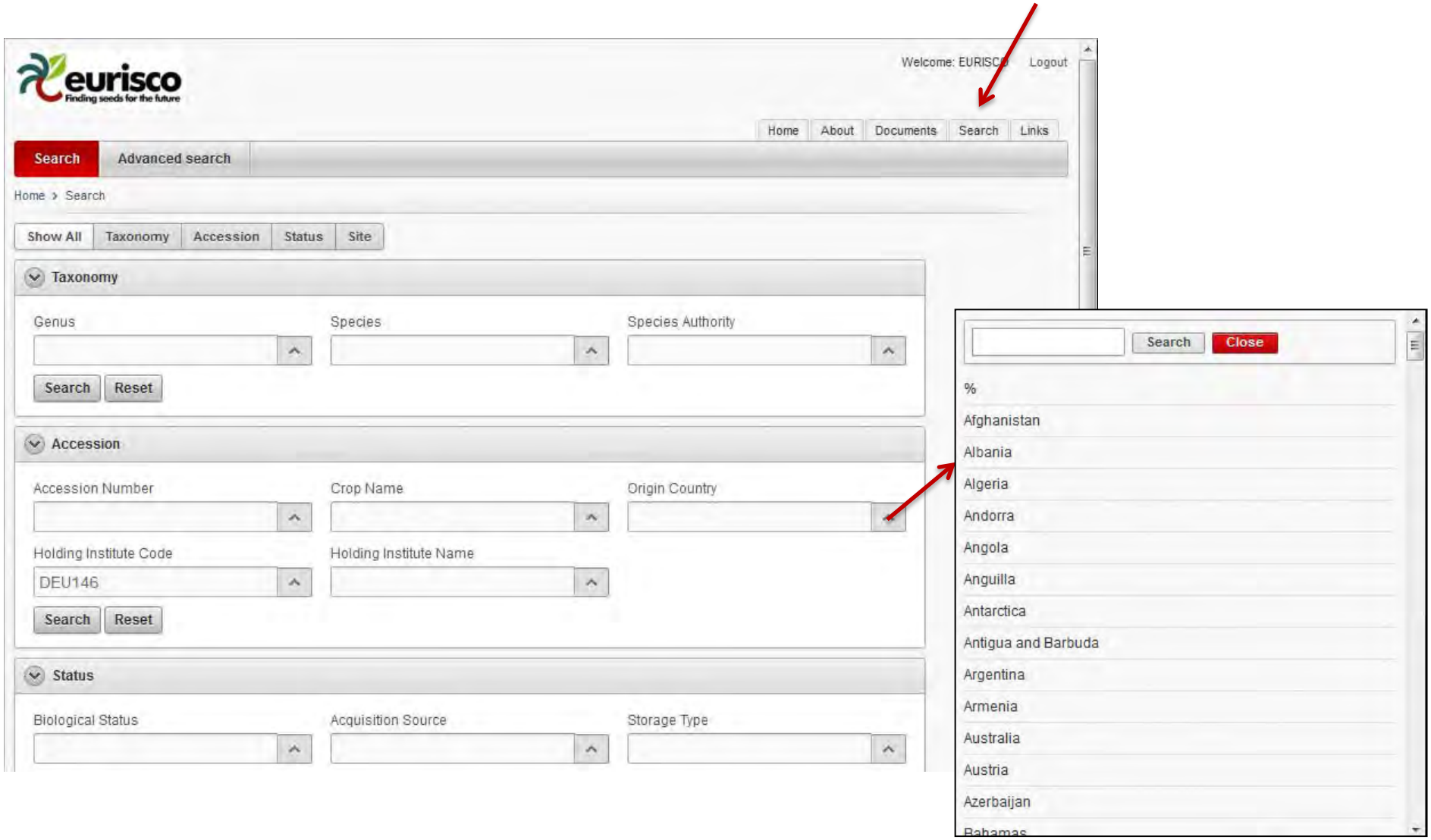
Reengineering V



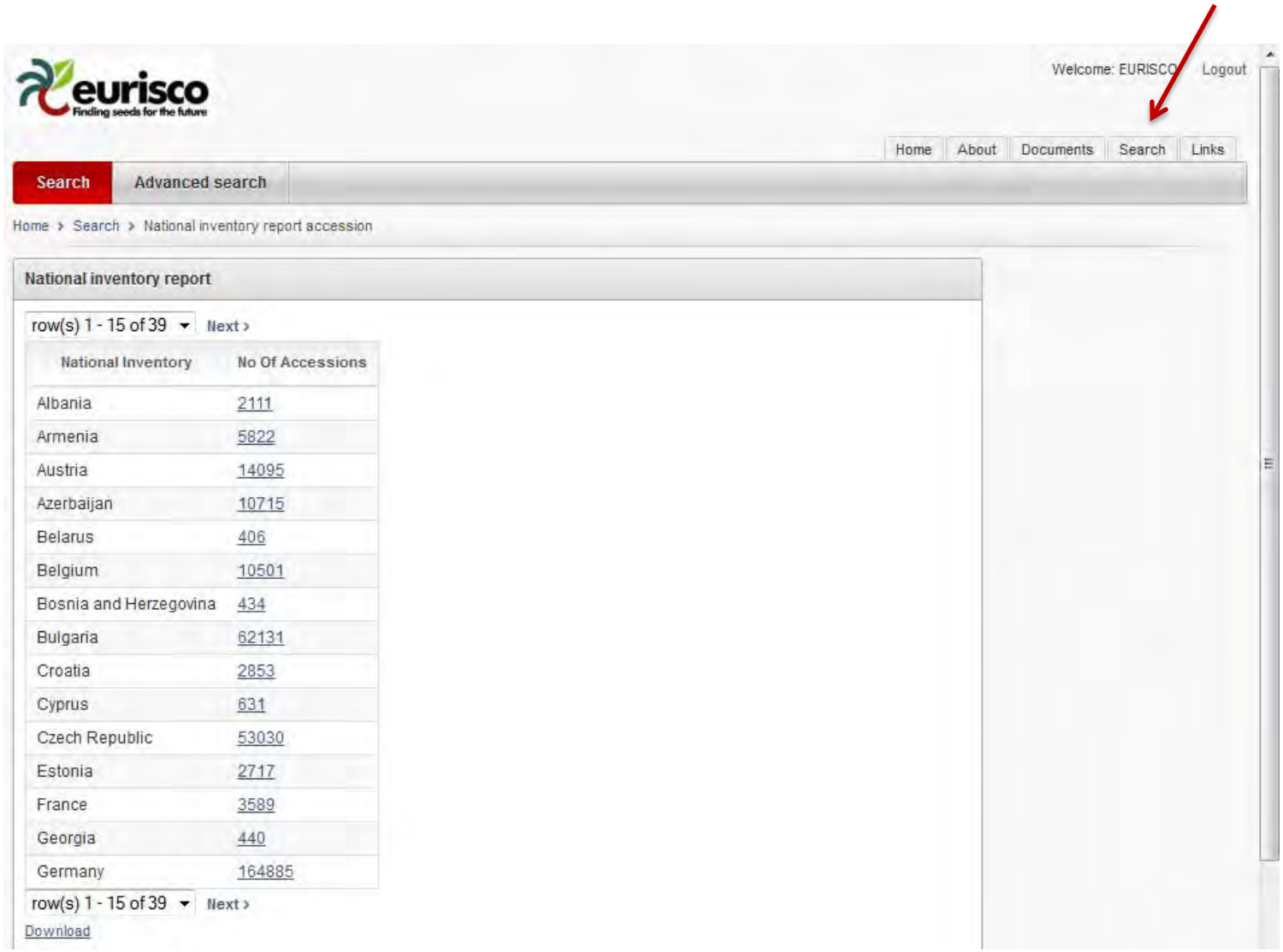
Reengineering VI



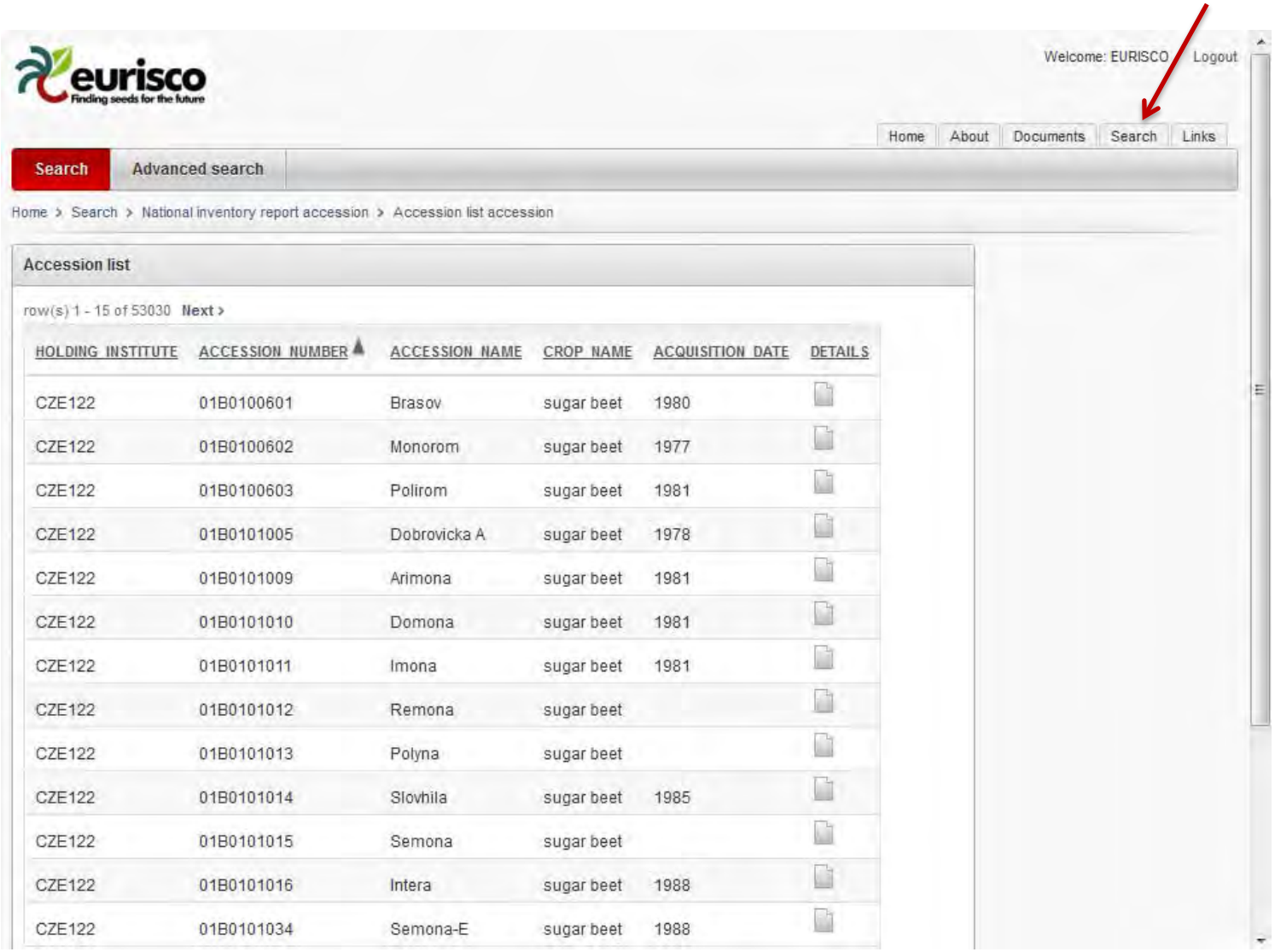
Reengineering VII



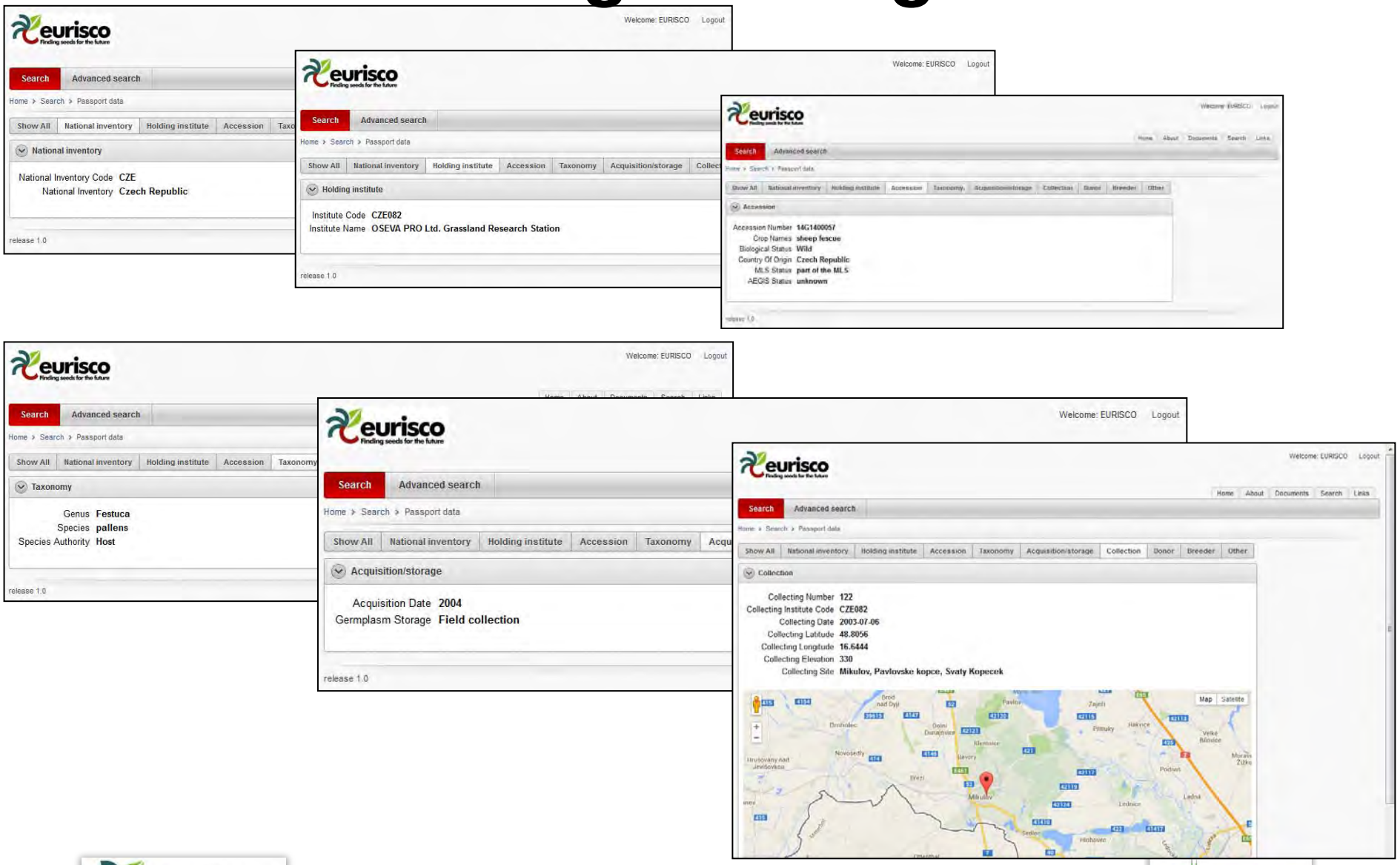
Reengineering VIII



Reengineering IX



Reengineering X



Reengineering XI

- To be done:
  - Development of new import component for NIs (former EURISCO intranet)
- Process will by monitored by IPK QM system
- EURISCO network
  - Focus on network activities after establishment of a working basis
  - Revitalise EURISCO network
  - Renewal of e-bulletins
  - Training workshops for data providers (NIs)

Outline

- Current status
- EURISCO transfer and reengineering
- Future developments

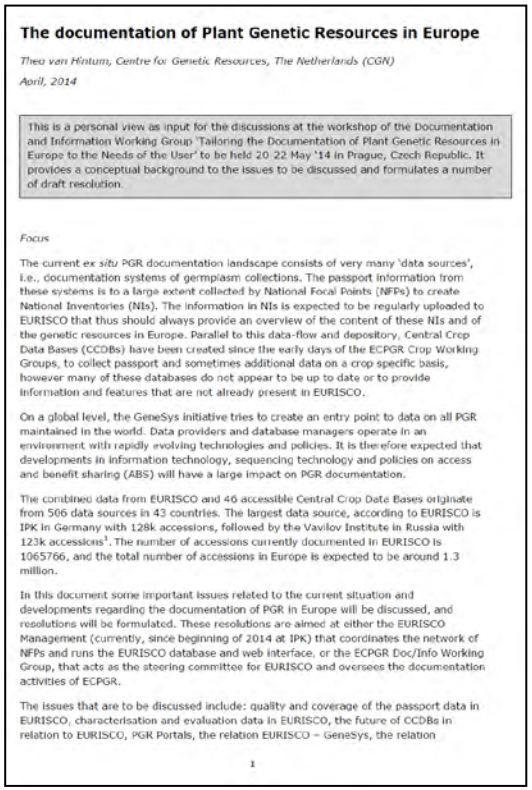
Future of EURISCO I

- Variety of challenges and ideas existing
  - Requirements from the PGR community
    - Especially from EPRGIS 3
  - Support for plant breeding and research
  - Further ideas

Future of EURISCO II

- Data quantity
  - Inclusion of additional passport data
    - ~250,000 accessions still missing (van Hintum, 2014)
- Data quality
  - Increase frequency of updates
    - Avg. age 1.16 years
    - Oldest 10% avg. 4.96 years

    (van Hintum, 2014)



The documentation of Plant Genetic Resources in Europe

Theo van Hintum, Centre for Genetic Resources, The Netherlands (CGN)

April, 2014

This is a personal view as input for the discussions at the workshop of the Documentation and Information Working Group 'Tailoring the Documentation of Plant Genetic Resources in Europe to the Needs of the User' to be held 20-22 May '14 in Prague, Czech Republic. It provides a conceptual background to the issues to be discussed and formulates a number of draft resolution.

Focus

The current *ex situ* PGR documentation landscape consists of very many 'data sources', i.e., documentation systems of germplasm collections. The passport information from these systems is to a large extent collected by National Focal Points (NFPs) to create National Inventories (NIs). The information in NIs is expected to be regularly uploaded to EURISCO that thus should always provide an overview of the content of these NIs and of the genetic resources in Europe. Parallel to this data-flow and depository, Central Crop Data Bases (CCDBs) have been created since the early days of the ECPGR Crop Working Groups, to collect passport and sometimes additional data on a crop specific basis, however many of these databases do not appear to be up to date or to provide information and features that are not already present in EURISCO.

On a global level, the GeneSys initiative tries to create an entry point to data on all PGR maintained in the world. Data providers and database managers operate in an environment with rapidly evolving technologies and policies. It is therefore expected that developments in information technology, sequencing technology and policies on access and benefit sharing (ABS) will have a large impact on PGR documentation.

The combined data from EURISCO and 46 accessible Central Crop Data Bases originate from 506 data sources in 43 countries. The largest data source, according to EURISCO is IPK in Germany with 128k accessions, followed by the Vavilov Institute in Russia with 123k accessions[1]. The number of accessions currently documented in EURISCO is 1065766, and the total number of accessions in Europe is expected to be around 1.3 million.

In this document some important issues related to the current situation and developments regarding the documentation of PGR in Europe will be discussed, and resolutions will be formulated. These resolutions are aimed at either the EURISCO Management (currently, since beginning of 2014 at IPK) that coordinates the network of NFPs and runs the EURISCO database and web interface, or the ECPGR Doc/Info Working Group, that acts as the steering committee for EURISCO and oversees the documentation activities of ECPGR.

The issues that are to be discussed include: quality and coverage of the passport data in EURISCO, characterisation and evaluation data in EURISCO, the future of CCDBs in relation to EURISCO, PGR Portals, the relation EURISCO – GeneSys, the relation

1

Future of EURISCO III

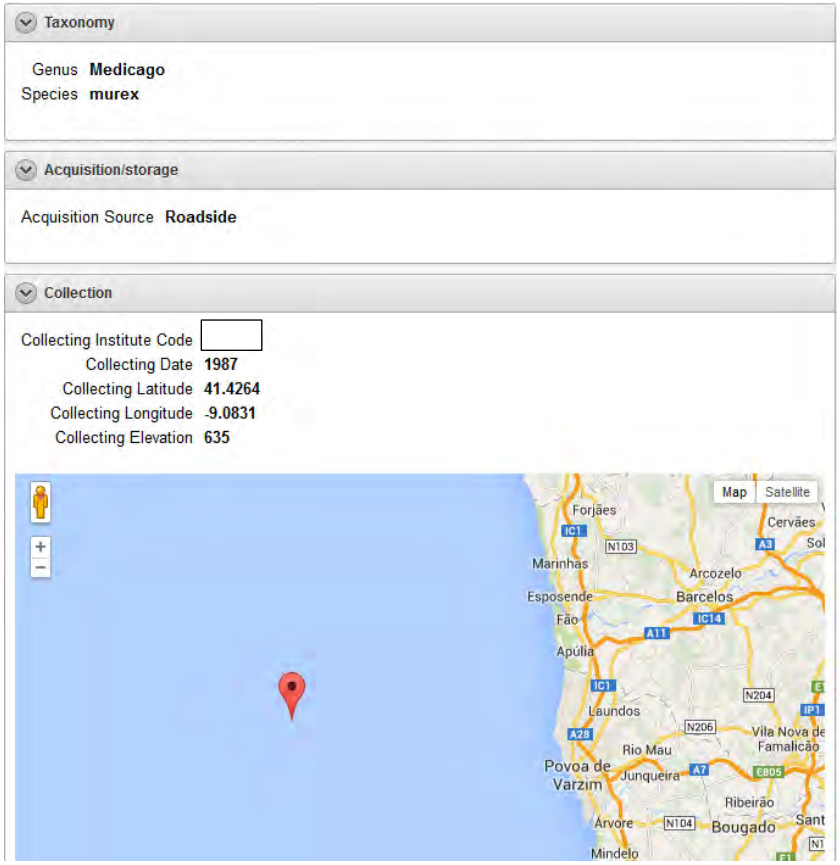
- Data quality
  - Improve taxonomic backbone of EURISCO
    - Management of taxon synonyms
    - Improvement of checks during import
      - GRIN, Catalogue of Life web service
    - Use of Taxonomic Serial Number?
      - Unique and persistent numeric ID
  - Increase completeness of information
    - Often limited information about certain accessions
    - Collecting information only for 39% of accessions

http://www.ars-grin.gov

http://www.catalogueoflife.org

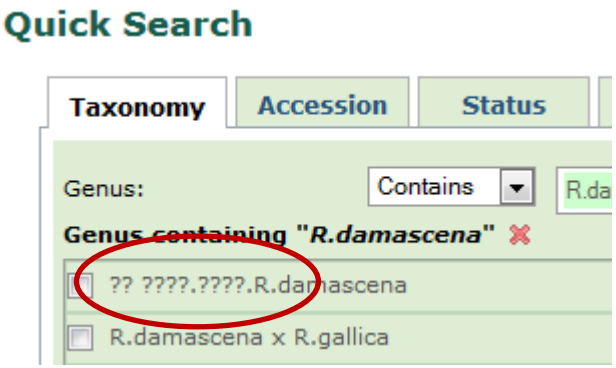
Future of EURISCO IV

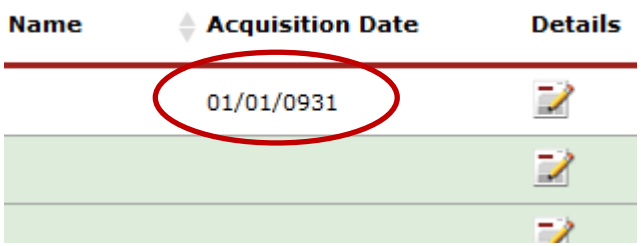
- Data quality
  - Improve location data quality

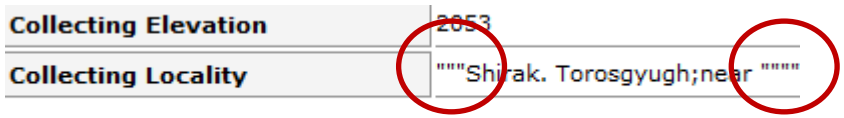


Future of EURISCO V

- Data quality
  - Reduce inconsistencies

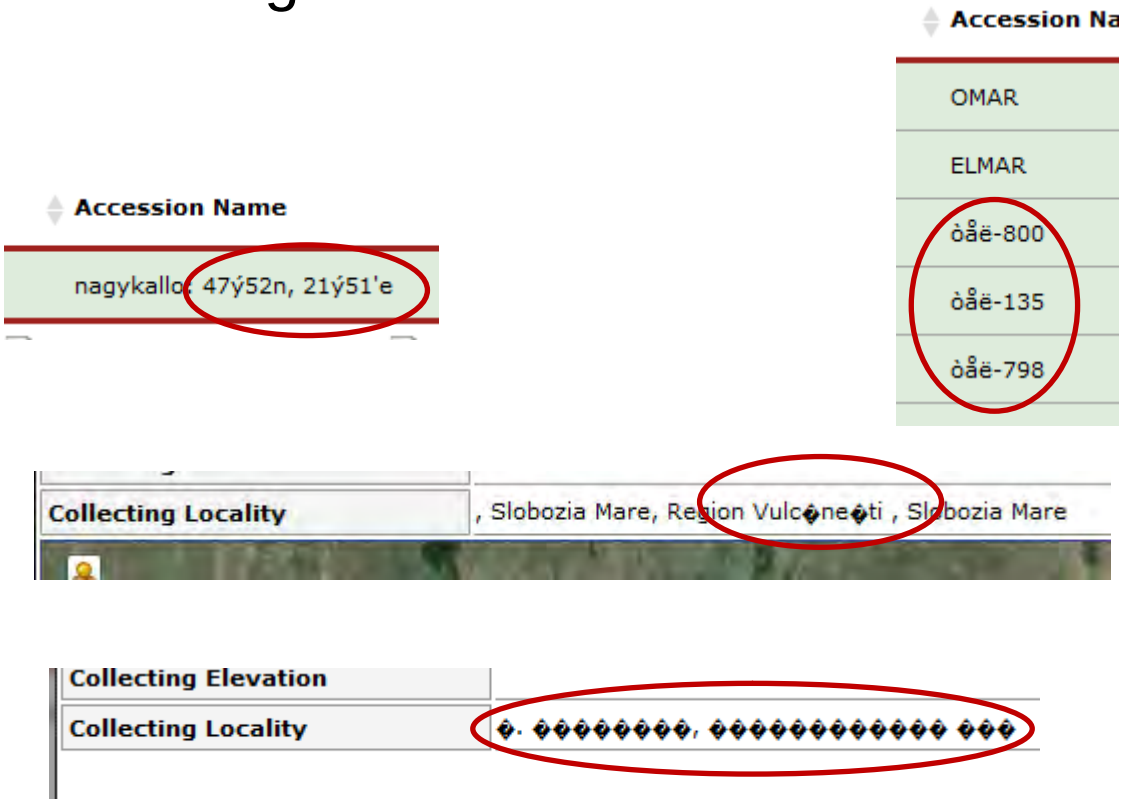






Future of EURISCO VI

- Data quality
  - Fix encoding issues



Future of EURISCO VII

- Improve import mechanism
  - Encoding issues
  - MS Excel files ↔ tab delimited files
    - Challenges with large data sets
  - Nested user accounts
    - Data upload by holding institutes
    - Check and approval by NIs

Future of EURISCO VIII

- Improve update mechanism
  - From full replacement to real update
  - Only incremental data needs to be updated
  - Necessary: Unique identifiers
    - Currently: Combination of NICODE, INSTCODE, ACCENUMB and GENUS
    - Possibility: Switch to real unique identifiers, e.g., LSIDs
    - Important for managing C&E data

Future of EURISCO IX

- Web services
  - Additional means of access
    - Data exchange with Genesys, GBIF etc.
  - Improvement of upload mechanism
    - Selective updates of certain accessions

Future of EURISCO X

- Support Central Crop Databases
  - Different quality
    - Well-maintained, incl. C&E data
    - Non-existing or many years old
  - In parts different acc. than in EURISCO
  - Instead: EURISCO as a central resource
    - DB managers could focus on crop-specific aspects, e.g., crop portals

Future of EURISCO XI

- Decision about *in situ* and on-farm data
  - Could be documented in EURISCO
  - Sufficient specification needed
  - Data exchange formats need to be agreed

- Make AEGIS status traceable
  - Audit flagging/de-flagging of accessions
  - Necessary: Unique identifiers

Future of EURISCO XII

- Develop EURISCO into “one-stop-shop” for germplasm requests
  - Allow users to select germplasm from multiple gene banks + submit seed request
  - Challenges
    - Information about availability indispensable
    - Up-to-date information needed
    - Automatic requests by email could generate high workload for genebanks
    - SMTA signature needs to be handled
    - Alternative: EURISCO could communicate with local ordering systems

Future of EURISCO XIII

- Inclusion of C&E data
  - Currently not available in EURISCO
  - Of high importance to users of PGR data
    - Determine value of germplasm for breeding and research
  - Difficult to handle due to lots of “standards”
    - Different descriptor names/synonyms
    - Different rating scales
      - Nominal, ordinal, metric scale
    - Different amounts of meta information
      - When, where, how, by whom?
      - Experiment set-up, treatment etc.

Future of EURISCO XIV

- Inclusion of C&E data
  - Pragmatic approach:
    - No standardisation of trait, scale or experimental design
    - Only standardisation of exchange format
      - As simple as possible
    - Import of existing data as-is
    - Additional:
      - Interpreted data could be displayed as added value
      - Several approaches existing
      - Input for CCDBs

Future of EURISCO XV

- Open EURISCO for additional C&E data
  - EU-funded projects or other collaborative projects dealing with gene bank material
  - Safe long-term maintenance of this data needed
  - EURISCO could be a repository for this data
  - Prerequisite:
    - Documentation of material in EURISCO (passport data) + C&E data public
    - Funding?

Future of EURISCO XVI

- Gene banks are more than “storage facilities”
  - Must be attractive for research and breeding
  - Aim: Inclusion of genetic information (“–omics”)
    - Improvement of collection management
      - Duplicates
      - Genetic purity
      - Core collections
    - Association studies with phenotypic data
    - Identification of appropriate germplasm for breeding
  - Obtain synergy effects from existing initiatives
    - Data management and visualisation

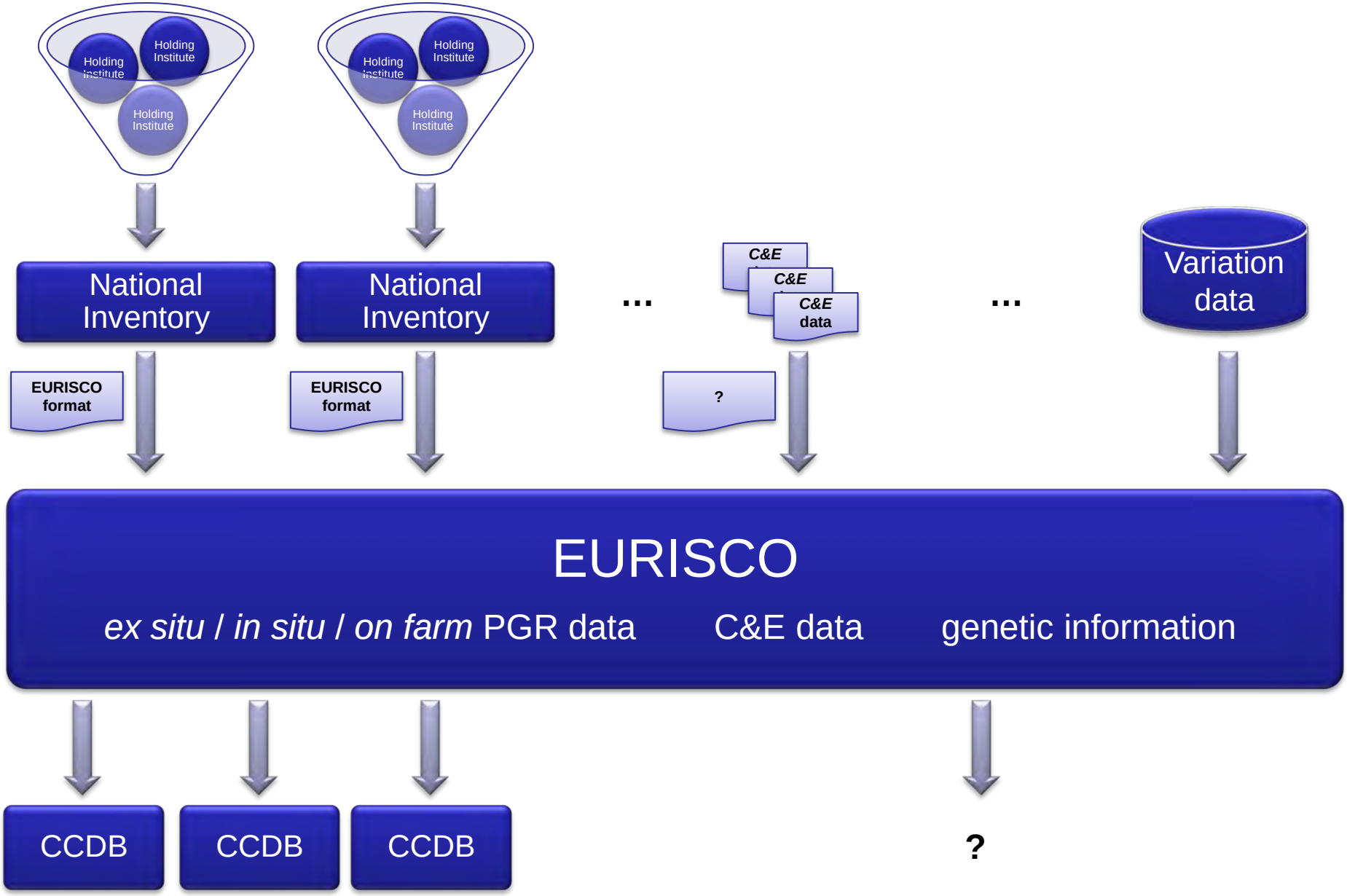
Proposed list of priorities I

- Short term
  - Finish consolidation of EURISCO
  - Revitalise EURISCO network
  - C&E data

- Medium term
  - Establish EURISCO as resource for CCDBs
  - Open for genetic information
  - Possible extension for *in situ*/on-farm data
  - Open for C&E data from EU-funded and other collaborative projects

Proposed list of priorities II

- Long term
  - Implementation of globally unique identifiers for accessions
  - Develop “one-stop-shop”


Holding Institute
Holding Institute
Holding Institute
Holding Institute
Holding Institute
Holding Institute
National Inventory
National Inventory
...
C&E data
...
Variation data
EURISCO format
EURISCO format
?
EURISCO
ex situ / in situ / on farm PGR data
C&E data
genetic information
CCDB
CCDB
CCDB
?

Acknowledgements

- Bioversity International
  - Sónia Dias
  - Simone Mori
  - Milko Škofič
  - Dario Valori
- Global Crop Diversity Trust
  - Matija Obreza
- IPK Gatersleben
  - Andreas Graner
  - Helmut Knüpffer
  - Markus Oppermann
  - Uwe Scholz
- ECPGR
  - Theo van Hintum (Doc&Info working group)
  - Gert Kleijer (Executive Committee)
  - Lorenzo Maggioni (ECPGR Secretariat)

EURISCO contact:

Stephan Weise

Email: weise@ipk-gatersleben.de, eurisco@ipk-gatersleben.de

Phone: +49 39482 5-744 Skype: weise_stephan