# WheatIS project

## *An integrated information system for the wheat research community*

**Date:** February 13, 2014

**Chair**: Hadi Quesneville

**Co-Chairs**: Mario Caccamo, David Edwards, Gerard Lazo

**Expert Working Group members:** Alaux Michael, Bastow Ruth, Baumann Ute, Clarke Fran, Dubcovsky Jorge, Edwards David, Herrero Javier, Itoh Takeshi, Kersey Paul, Marshall David, Martinez Cesar, Matthews Dave, Mayer Klaus, N'Diaye Amidou, Rawlings Christopher, Franck Röber, Doreen Ware.

## Introduction

This project aims at building an International Wheat Information System, called hereafter WheatIS, to support the wheat research community. The main objective is to provide a single-access web base system to access to the available data resources and bioinformatics tools.

This project is based on the principles listed below:

- Collective building of the WheatIS to better respond to the needs of the international wheat community;
- Incremental implementation to offer rapidly an operational information system;
- Emphasis on Quality Assurance to serve as a framework for an approach with incremental implementation;
- Promotion of an open-access model for data exchange;
- Reliance on a distributed system;
- Use of Virtual Machine and *Cloud Computing* technologies to facilitate sharing data and tools;
- Promotion of the visibility of each participating platform to contribute to their sustainability.

## WheatIS Expert Working Group (EWG)

The project is driven by the wheatIS EWG, a network of 21 experts that congregates a group of volunteers willing to participate to the WheatIS project, selected after an open call for proposals. They elect for a 2 years period a chair and 3 co-chairs to lead the group. One co-chair is changed every year.

The deliverables of the WheatIS EWG will be:

1- Creation and animation of a network of experts, collaborating to provide the scientific community with wheat genetic and genomic data.

2- Standards, protocols, and processes for wheat data sharing in keeping with different international bioinformatics initiatives. Guidelines and recommendations will be disseminated and shared between bioinformatics teams and the scientific community. These documents will be accessed through a web site and regularly updated according to the evolution of the scientific fields.

3- A web platform allowing the exchange of standardized data files, built on a collaborative and interoperable network of platforms, working together.

4- A single entry point for the wheat community to find available data through a full text search engine, allowing searching the central repository and the databases of the platform network dynamically.

5- Integrated data focusing on relevant data sets, chosen with the wheat scientific community.

6- Develop Project work plan and annual activity reports of the EWG activities.

7- Research proposals to get funds for the implementation of the WheatIS.

## WheatIS overview

The WheatIS will operate as a hub integrating wheat data produced and submitted to the public repositories by the community. It will rely on a network of bioinformatic platforms willing to contribute and to work in synergy to provide the wheat scientific community an easy access to wheat data. These platforms, each being considered as a WheatIS node (Figure 1), will share their resources (staff, storage, infrastructure, and tools).

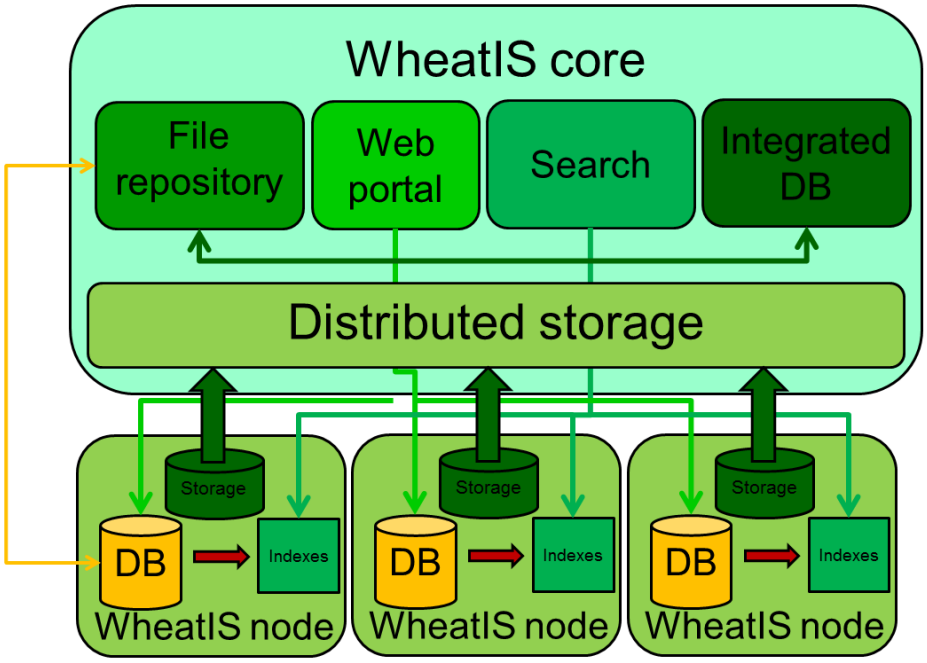


Figure 1: WheatIS general architecture

A central node, called WheatIS core, will provide a single entry point for the WheatIS users. The WheatIS core will be built upon the resources provided and shared by the nodes. It will provide access to data and information through a web portal. This portal will give access to a data file repository storing files with their associated metadata. A google-like search engine will allow to find data available in the WheatIS core and its nodes using keywords. Several dedicated integrative databases, e.g. for genomic, genetic, and phenotype information, comparative genomics, and functional genomics, will be available. Integration with environmental data will be considered when available.

Analysis tools will be available for download from the web portal. Some WheatIS nodes will provide computing resources for data analysis.

# Project framework

## Rationale

Instead of proposing a static view of the WheatIS, we present a dynamic model that will be adapted as the system implementation progresses, and where each step builds optimally on the previous one. The rationale is that considering the complexity of building the WheatIS, it will be preferable to build it progressively, learning from each phase by the successes and the failures and correcting the strategy accordingly. Tools and technologies will undoubtedly evolve rapidly in the next coming years. Users' needs will be considered as well, as they will evolve in parallel. This iterative process will have a better chance to succeed, as its evolution will benefit from the community feedback at each step.

WheatIS EWG members will be involved in other international bioinformatics initiatives (such as Ontology, International Arabidopsis Informatics Consortium, TransPLANT) to develop synergies and collaborations.

The WheatIS project is divided in 6 work packages (Figure 2), each focusing on specific objectives and tasks. The content of each work package (WP) will be detailed and enriched by the WheatIS EWG when set up. The deliverables of each WP will be released at first in their simplest, most easily achieved form to provide rapidly operational solutions to the scientific community, but will be used to build more elaborate ones iteratively.

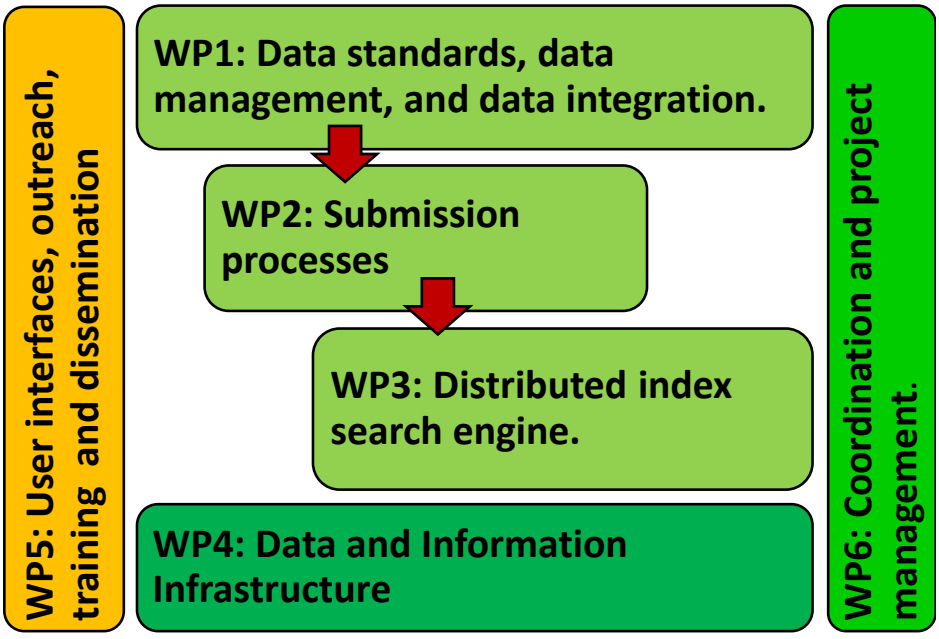


Figure 2: Project overview

## WP1: Data standards, data management, and data integration

### Objectives

This work package will define standards for data exchanges. Part of the activity will be to follow and participate in international initiatives (e.g. TransPlant, crop and trait Ontologies, etc.) and to propose standards when they are missing.

Coordination and implementation of data integration is also a goal of this work-package. It will rely on free data exchanges between WheatIS nodes to integrate differently the same data according to different users' needs. Hence, several WheatIS nodes for data integration will be identified. Data curation managed by WheatIS nodes will insure reliable data integration.

This work package will interact closely with WP5: User interfaces, outreach, training and dissemination.

### Tasks

- Survey of existing standards
- Coordination of data exchanges and data release policy
- Identification of end-users' categories and WheatIS nodes to support their activities
- Data curation management
- Implementation of Web services and database mediation (DAS, InterMine, BioMart).

## WP2: Data submission process

### Objectives

This work package will provide user support for data submission to the WheatIS or other international repositories (NCBI, EMBL, etc.).

The WheatIS will offer a space allowing data submission and the exchange of standardized data files with their associated metadata. The submission process will check for required metadata information and the correct use of data standards and ontologies as defined by WP1.

This space will be also used for data exchange between WheatIS nodes or any wheat scientists. Data correctly formatted and respecting defined standards will be provided here promoting data interoperability.

### Tasks

- Implement a distributed data file repository with web access and data submission workflow. Test available tools such as Dspace
- Ensure scalability and security through data replication and data sharing. Test existing solutions such as the integrated Rule Oriented Data System (iRODS)

## WP2: Distributed index search engine

### Objectives

A full text search engine will be set up on the WheatIS web portal, allowing to dynamically search remote WheatIS nodes databases (see figure 1). Users will be able to connect to the WheatIS portal and type a keyword, or a term, that will be searched remotely in each node. Results will be provided as a brief summary of the matching data (e.g. Identifier, Name, Short description) with links to access the remotely hosted data. This tool will allow the researchers to discover available data in the WheatIS core and its nodes, from simple keywords. Indexing with public Google web and Google scholar search engines will allow scientists to find data through the Google interface without necessarily knowing the WheatIS portal, hence offering another way to discover its existence.

### Tasks

- Implement quick search functionality using distributed indexes. Implement SolR servers in the WheatIS nodes
- Define file formats for data extracted from databases
- Implement a standalone server able to convert extracted files into searchable indexes

## WP4: Data and information infrastructure

### Objectives

Distributed architecture is the backbone of the WheatIS. It will use Data Grid technologies implemented mainly on virtual machines to easily distribute portal instances between WheatIS nodes, but also to insure robustness, speed, and easy development.

### Tasks

- Identify requirements and design integrating architecture
- Virtual Machine implementation
- Hardware infrastructure setup
- Building of Cloud environment

## WP5: User interfaces, outreach, training and dissemination

### Objectives

This work package will ensure that the WheatIS project is implemented in interaction with users and other international bioinformatics initiatives. It will establish links with the wheat research community to inform them on the progresses, get their feedback and answer their needs.

Hence, integrated activities with the International Wheat Genome Sequencing Consortium (IWGSC) EWG will be developed to define gene annotation nomenclature and annotation, and reference sequence life cycle. The WheatIS will need to be sufficiently developed in 2016 to display and share the data produced by the IWGSC (reference sequence, annotation, etc.). Agreement for data access and data release policy will be supported by the WheatIS. Tight connections with other EWG will be

maintained, when they are created, to insure proper data sharing and integration for the various topics of interest.

A web portal will be implemented and maintained to provide a single access point to the WheatIS data, the bioinformatics services offered by the nodes, informations on the project, and guidelines and good practices for data sharing and data submission.

### Tasks

- Build and maintain the WheatIS web portal
- Communicate through the web portal, mailing lists, twitter, facebook, etc.
- Organise users' survey to get their needs and measure the satisfaction
- Organise bioinformatics training for the wheat research community
- Create or strengthen links with international initiatives related to wheat research or plant bioinformatics
- Attend international conferences and meetings, present WheatIS results and report other groups' advances and needs

## WP6: Coordination and project management

### Objectives

This work package manages, coordinates and communicates on the EWG activities and the WheatIS project.

During the project duration, annual meetings will be organised to discuss the progress, the needs and the future orientations. They will be organised in coordination with other international meetings to facilitate logistics and limit travels cost. Regular teleconference calls (at least 3 per year) will maintain the contact between the participants. Working groups will interact monthly to work on specific topics. Annual activity reports will be provided to the Wheat Initiative committees.

### Tasks

- Set up governance - organise committees and election of chairs
- Provide a project workspace to exchange documents between EWG members
- Organise meetings
- Write annual reports

# Implementation

We propose to build the WheatIS in three main steps starting from a low-technology, easy-to-achieve infrastructure, towards a more ambitious integrated system.

- Step 1 : Network building (year 2013 to be continued in 2014)

- Step 2 : Integrated virtual portal (year 2014-2015)
- Step 3 : Integrated database (year 2016)

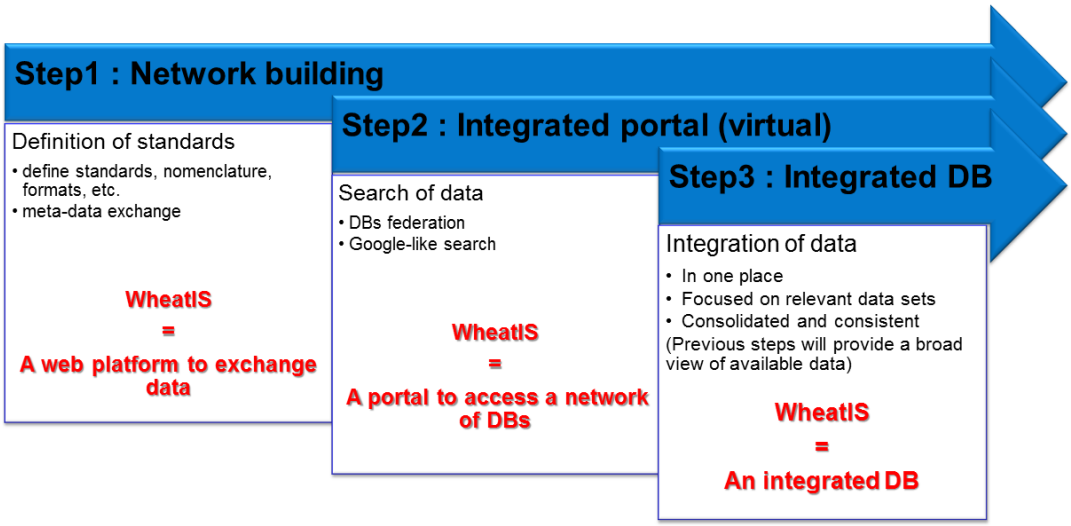


*Figure 2: WheatIS timeline overview*

Step 1 and 2 will provide a broad view of the available data. Step 3 will integrate the data in one single, centralised information system. Being a major task, integration will focus on relevant data sets, chosen with the wheat scientific community. It will not replace the tools built in steps 1 and 2, but will add a new browsing functionality, allowing users to navigate through data and explore their relationships. It will also answer complex queries involving data hosted initially in different locations (files, databases). This system will also be able to produce integrated, consolidated and consistent information, which could be exported as data files to feed analysis pipelines or other information systems.

For year 1 (2013), we plan to:

- Organize the first WheatIS EWG meeting
- Write a detailed plan of activities
- Submit a research proposal to get funds for the implementation of the WheatIS

## Funding the project

Building and maintenance of the WheatIS will rely primarily on the contributions of the partner WheatIS nodes and their funding institutions, each of them dedicating resources to the project. Additional resources will be looked for from national or international calls for proposals, with the EWG coordinating proposals and writing letters of support.

Steps will be implemented rapidly by distributing solutions available in a WheatIS node partner to the other partners, in a transparent way for the end-users. This approach would facilitate other WheatIS nodes (not yet identified) to join the initiative. This will however necessitate the adoption by all platforms of solutions provided by others.

The project will start from a group of WheatIS nodes having received funds from their supporting institutions. WP1, 5 and 6 could start with no additional resources. However, for a quick start of this project, WP2, 3 and 4 will require newly recruited staff paid with institutional funds. Infrastructure capacities should be supported by the institutions directly to their WheatIS node.

One coordination meeting between the WheatIS nodes will be organised each year (approximately 50,000 euros per year including travels and accommodation for 30 persons), which should be supported by the Wheat Initiative in the absence of other funding.