Towards the construction of an integrated Wheat Information System

Mario Caccamo, Hadi Quesneville

Towards the construction of an integrated Wheat Information System

Report-June 2012

Mario Caccamo [1], Hadi Quesneville [2]

1. The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, UK
2. INRA, Research Unit in Genomics-Info, Versailles, France

Introduction

This report was commissioned by the *Wheat Initiative* scientific board following the proposal adopted by the G20 Agriculture Ministers in June 2011. The aim of this report is to offer recommendations on the best strategies to follow to develop an integrated Wheat Information System (hereafter called WheatIS) and provide the international wheat research community easy access to wheat genetic information, genomic data and bioinformatics tools. This report is based on results obtained from a community-wide consultation and opens the possibility of integrating agronomic data within the WheatIS.

This report was built in collaboration with an expert committee composed of wheat scientists and other researchers involved in the development of information systems initiatives (see members names in appendix 1).

I. Community-wide consultation results

A web-based survey questionnaire was implemented to consult the wheat research community on areas of interest to them. The link was circulated to recipients through several mailing lists and relayed by scientists to their collaborators. We also collected feedback from a dedicated workshop that we organised with the support of the Wheat Initiative Scientific Board during the 20th Plant and Animal Genome meeting, held in San Diego in January 2012.

Web-survey

The questionnaire for the consultation was prepared with the input from the expert committee and implemented as web pages to be filled by the survey participants. The survey was open from 12/14/11 to the 03/14/12 and was answered by 282 participants from 30 different countries (Figure 1).

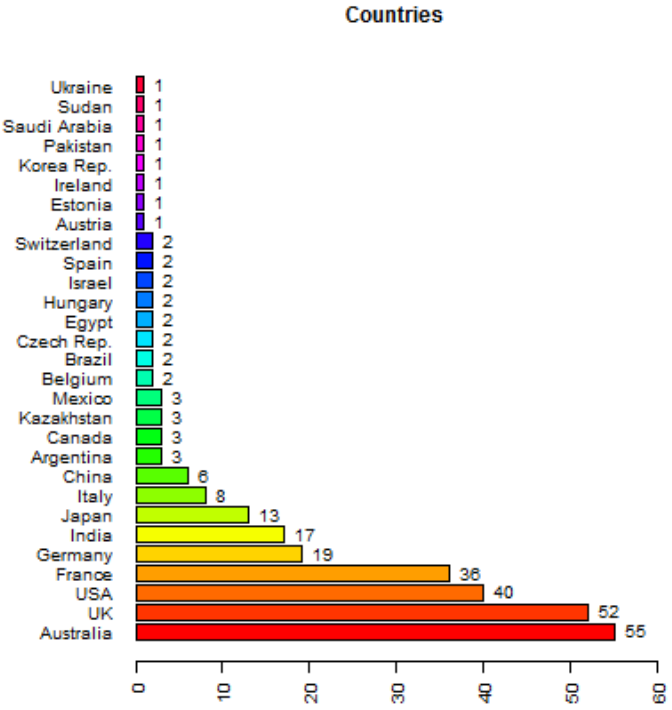

Figure1: *Number of respondents per country*

The fields of expertise of the surveyed scientists covered the WheatIS targeted scientific domains (Figure 2) Breeding and functional genetics were the most represented.

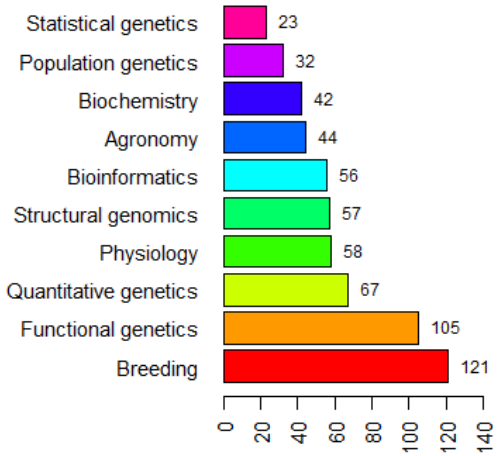

Figure 2*: Breakdown of areas of expertise of respondents (number of respondents per field of expertise)*

Asked the type of analysis they would like the WheatIS to support, wheat researchers ranked functional and genotype to phenotype analysis ahead of comparative genomics and breeding (Figure 3).

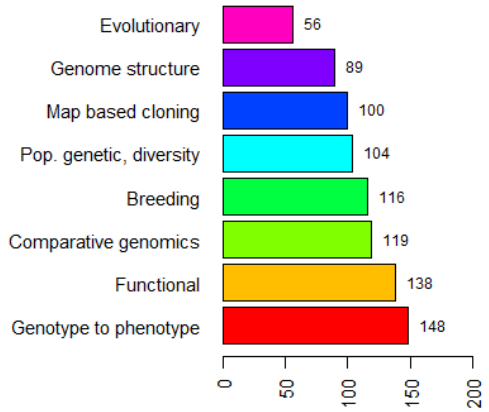

Figure 3*: Preferences for analysis types, ranked by responses*

To help with develop and populate the WheatIS, with the many types of data that wheat researchers will need to access, we were interested in which data researchers viewed as a priority. Accordingly, we asked which data appeared to be the most important for wheat research in the coming 5 years. When sorted by importance (Figure 4), we observed a smooth progression in the type of data. Although there is not a clear-cut preference, SNPs, genome assemblies, phenotypes, maps and molecular markers were marginal leaders, which may reflect a current trend in technologies and resources building in the community. However, all data types could be considered important for the future development of wheat research.

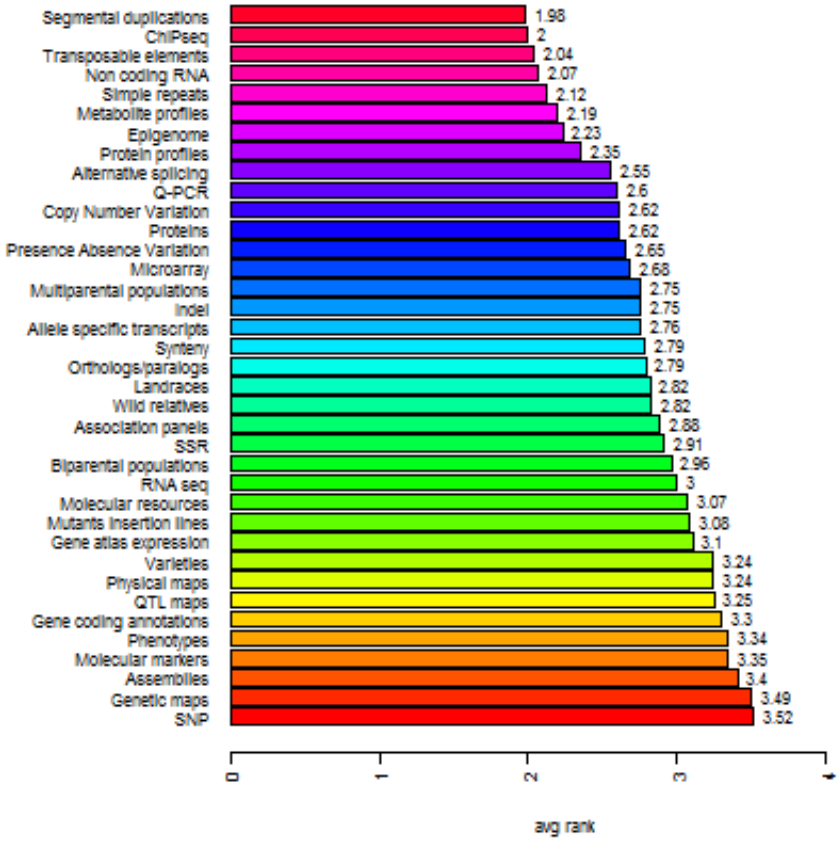

Figure 4: *Average rank of importance of wheat data types in the coming five years*

The WheatIS could operate as a hub and integrate wheat data produced and submitted to the public repositories by the community. In combination with it, data could be transferred and stored on other bioinformatics portals. When asked on which existing platform they would prefer to also store their data, wheat researchers placed the NCBI and GrainGenes portals in first and second position respectively, followed by Gramene and Wheatgenome.info (Figure 5A). However, marked differences were observed between countries (Figure 5B). Considering the 4 countries most represented in the survey, GrainGenes and NCBI were preferred by researchers based in Australia, GrainGenes in the USA, NCBI and Cerealsdb in the UK and URGI in France. This might reflect preferences related to users considerations or specific research interests, and/or the knowledge of local databases by national research communities.

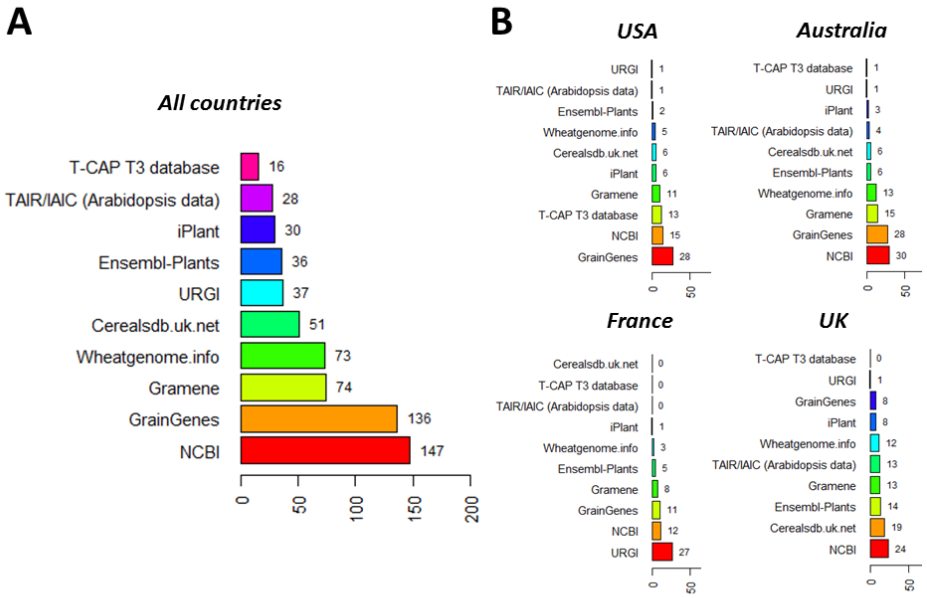

Figure 5*: Other bioinformatics portals where data would have to be accessible from (X axis: number of positive answers)*

When asked what kind of services the WheatIS should provide, wheat researchers placed data browsing, data downloading and genome viewer respectively in first, second and third position (Figure 6). Interestingly, data integration was placed only in fourth position, indicating that this is not considered yet as a top priority. Surprisingly, we also noted that analysis workflows and computing capacities were ranked last. Again this might be due to current technological limitations and lack of access to more advanced bioinformatics tools.

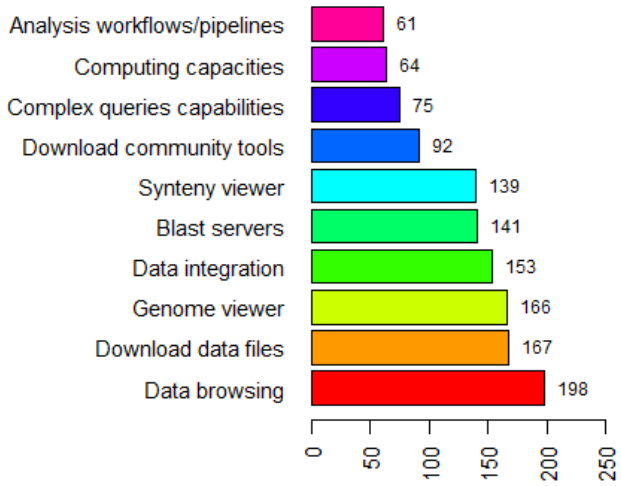

Figure 6*: Preferences for services, ranked by number of responses*

Most of the participants supported the data release policy developed by the Bermuda / Fort Lauderdale / Toronto agreements (Nature 461, 168-170, doi:10.1038/461168a), that promotes the early dissemination of whole-genome datasets but preserves the rights for the data generators to lead the analysis and publication of their data in peer-reviewed journals (Figure 7). Nonetheless, a substantial number of responses were also in favour of a 1 to 6-month period embargo.

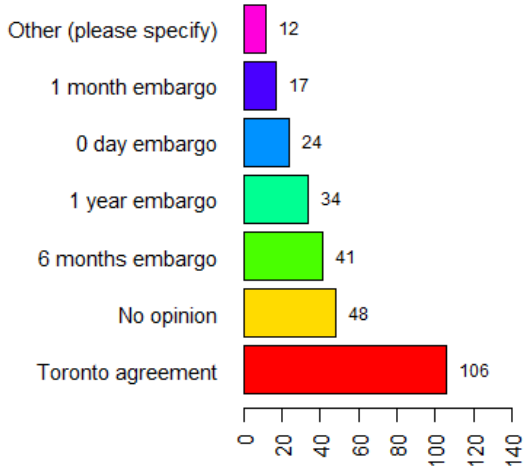

Figure 7*: Preferences for data release policy, ranked by number of responses*

Users' views would have to be collected regularly in order to ensure that the WheatIS continues to meet the community's needs. The survey indicated that web-surveys and users' committees composed by representative end-users were the preferred mechanisms to get the research community feedback (Figure 8).

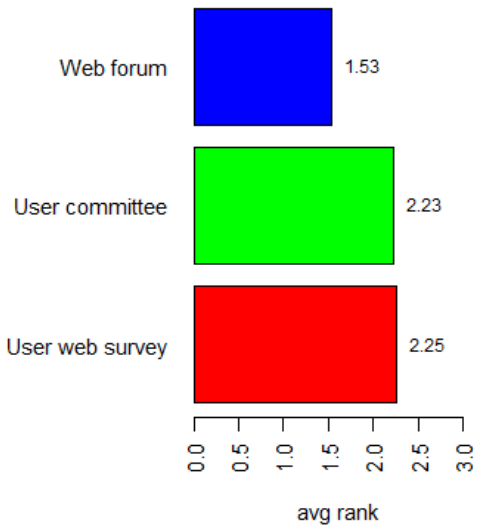

Figure 8: *Average rank of importance of users' needs survey mechanisms*

Plant and Animal Genomes (PAG XX) Conference Workshop

We organized a workshop open to the wheat scientific community during the 20th Plant and Animal Genomes (PAG) meeting held in San Diego in January 2012. We presented preliminary results of the web-survey and proposed a strategy to develop the WheatIS. Around 40 persons attended this event for a round-table discussion that lasted 2 hours.

Some highlights of the audience questions and comments are reported below:

- Data quality with respect to standards (protocol, nomenclature, format) is a key issue for data management. It also facilitates data diffusion to the community. The implementation of strategies for curation and quality checking will be an important aspect that the WheatIS should support. Legacy data are also important to consider as they could be used as gold standards for validation and comparisons. Hence, the added value of the WheatIS will be to provide sets of curated, formatted, and integrated data to other databases (wheat or other related species).
- Metadata should include information about the data providers so that their work can be acknowledged, but also to encourage data submission by giving visibility to the scientists who generated them. Literature related to a data set should also be included. Journals' policies to deposit the data in the WheatIS prior to publication should be encouraged by the Wheat Initiative consortium.
- The industry partners should be consulted to build a data management policy compatible with their requirements. It is highly desirable that they contribute as data providers, specially for pre-competitive data that could be beneficial for the global research community.
- Data integration should be driven by real case studies to avoid effort dispersion.
- Data should be accessible from facilities equipped with computing power and storage (e.g. cloud-based approaches) to facilitate running computing-intensive analyses.
- Multi-lingual information system could be important for breeders not familiar with English (e.g. researchers in emerging economies in Asia and South America).

- The audience pointed that haplotype/map data were not explicitly mentioned in the survey. They should be considered.
- The WheatIS should be built on experiences gained by similar initiatives (plant species, animals, and biomedical field). It will have to participate to the data management knowhow exchanges between bioinformatics platforms.

II. A proposal for the WheatIS architecture

This WheatIS proposal takes into account feedback from the wheat scientific community, from experts working on similar initiatives for other species and from the bioinformatics community in general. We built this proposal by taking into account the considerations given below.

First, the WheatIS will have to add value and represent a substantial improvement to what is already available for the community. It should not replace the current **wheat information databases**, but should offer them new services, for instance to tackle concrete data management issues such as curation, formatting and integration.

Second, several **bioinformatics platforms** dealing with wheat data already exist and provide useful services to the community. The WheatIS should be built on these platforms as a federated network of existing services, considering that:

i. Experiences from other species indicate a tendency to build such networks, instead of single information system (e.g. TAIR/IAIC, EBI/Elixir).
ii. Exiting platforms already offers some services to particular wheat communities. They are specialized on scientific fields, large projects, or countries. They have already served their community for several years, often doing a good job, as indicated by the web-survey. Consequently they are well positioned to continue in a larger framework.
iii. Providing a single information system which meets everyone's needs may be overambitious and hamper innovation and adaptability. The data are diverse and the number of scientific fields using them important. It might therefore be difficult to build a single platform in one place. We should rely on a world-wide network of expertise. Integration and standardization are important goals. Our aim should be to integrate searches across diverse databases and encourage format sharing.
iv. No dedicated funding by the Wheat Initiative is expected.

Hence, we propose that the WheatIS should be considered as a framework where data and expertise exchanges are facilitated and enhanced. In this framework, platforms will easily exchange their data to provide their users with up-to-date top-quality information. Users will also have access to enriched integrated and curated data from a central repository. With this in mind, we built a proposal consistent with this view from the infrastructure and governance perspectives.

An evolving infrastructure

Instead of proposing a static view of the infrastructure, we present a dynamic model that will be adapted as the system implementation progresses, and where each step builds optimally on the previous infrastructure. The rationale is that bearing in mind the complexity of building a WheatIS, it will be preferable to build it progressively, learning from each phase by the successes and the failures

and correcting the strategy accordingly. Users needs will be followed 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.

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

Step1 : Network building

The first step will be to build a collaborative and interoperable network of platforms, working together to set up the first visible WheatIS service. Initial tasks will be to define standards, formats, and nomenclatures. Data hosted by the WheatIS will follow the defined rules allowing homogeneity, coherence, and reliability between data. Hence, standardized data could be exchanged between users and easily reused in different analysis. Wheat IS partners will define standards collaboratively, taking into account current standardization works of different platforms. Data using this format will be submitted by scientists to a centralized web file repository. The WheatIS platforms will be able to help the community in this task. Success will depend on this being instigated by the Wheat Initiative consortium. In particular, agreements with peer-reviewed journals for proper data deposit in the centralized system will be key for the success of this initial step.

At this stage, the WheatIS will be a web platform allowing the exchange of standardized data files. It will be possible to search the WheatIS metadata using keywords or full text searches. Indexation through Google web and Google scholar search engines will be used to allow scientist to find data through the Google interface without necessarily knowing the WheatIS portal. Solutions are already available to implement such a web platform (e.g. dspace: www.dspace.org, iRODS: https://www.irods.org). Consequently, we consider this as an achievable and relatively low-tech solution.

In addition, WheatIS partners will be able to download data from the WheatIS repository into their own information system. This will allow users to view the WheatIS data alongside their own data, or to offer a specific service to a dedicated community (e.g. partners of a project). Indeed, several ways of integrating data could be proposed, each resulting in a specific integrated view, depending on the requirements of particular scientific fields (i.e. breeders, population geneticists, functional genomicists) allowing the same data to be viewed differently according to need. The data standardization and the central repository will greatly help the WheatIS partner platforms in this task.

Step2 : Integrated virtual portal

At the end of Step 1, the WheatIS will be seen as a network of platforms, sharing data files on a central web file repository, but each still maintaining their local databases. The second step will be to set up a full text search engine on the WheatIS portal, allowing to dynamically search these local platform databases. The users will be able to connect to the WheatIS portal and type a keyword or a term that will be searched remotely in each database. 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.

Step3 : Integrated database

The final step will be to integrate the data in one single, centralised information system. Previous steps will help to provide a broad view of the available data. 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 relationship. It will also answer complex queries involving data that hosted initially in different location (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.

Tools and technologies will undoubtedly evolve rapidly in the next coming years. The strategy presented here will be re-evaluated at each step and the WheatIS development strategy adapted accordingly. Agronomic data should be easily integrated in the system at a later stage. This will require dedicated bioinformatics platforms joining the WheatIS network. In addition, we recommend that the WheatIS follows the work of other international bioinformatics initiatives (such as Ontology, IAIC, TransPLANT) by developing synergies and collaborations.

Wheat research community needs: today and in five years.

According to the web-survey results, the main needs for the five coming years are focused on breeding and functional studies requiring information on genome sequence variants (SNP, Indels, CNV), genome assemblies, phenotypes, maps and molecular markers. Therefore, data management efforts should first concern these types of data.

Interactions with users

The web-survey analysis shows a large consensus in favour of a data release policy under the Bermuda / Fort Lauderdale / Toronto agreements. An unrestricted access will be offered to published data and most legacy data. Unpublished data will be distributed upon the signature of an agreement stating the respect of the data producers and the contributors rights to analyse and publish analyses in peer-reviewed publications. Data access could be managed through private accounts in the information system. Access rights will be granted according to the signature of a document. Note that data with restricted access could be accepted as it could also speed up data diffusion allowing submitters and WheatIS staff to check the data as it appears in the system before granting data access to others. However, this embargo duration should not last more than 6 months. Given the popularity of immediate data dissemination, data release under a Bermuda / Fort Lauderdale / Toronto agreement will be encouraged.

Users will have the opportunity to interact with WheatIS management regularly to give feedback and express their needs. Web-surveys will be sent to the community each year to get satisfaction feedback. A users' committee composed of end-user representatives will be set up to discuss the evolution of the system with the WheatIS steering committee. Satellite meetings will be organised around the main wheat conferences to present novelties and to discuss service improvements.

An e-mail help-desk will be available to help users and answer their problems.

Required resources

With no specific funding, the WheatIS will rely mainly on existing infrastructures and the contribution from the scientists and bioinformaticians from the community (e.g from wheat platforms that will directly participate in this effort). However, resources will be needed for the implementation of a central repository that should probably be hosted by one of the existing platforms. A key prerequisite

will be to guarantee the sustainability of this platform, and therefore the long-term availability of the following resources:

- Hardware infrastructure: petabyte-scale storage must be available with support for backups (such as mirror servers). Note that raw and curated data submitted to NCBI or EBI will be transferred regularly. Servers must be securely kept in a dedicated computer room with fire and intrusion protection.
- Personnel: Highly qualified staff will be required. The infrastructure will need database managers, system and network managers, developers and data managers. Around 10 full time engineers/researchers should work for the WheatIS.
- Quality control procedures: A management quality system must guarantee the correct functioning of the hosting platform. It will insure effectiveness of the procedures and protocols used the availability of the computer infrastructure, and the quality of the services.
- Financial sustainability: Long term service preservation and data accessibility must be guaranteed. Staff salaries, computer licences and maintenance fees, hardware renewal costs, travel and accommodation expenses to coordinating and scientific meetings must be covered.

Governance

The proposed governance is composed of a users committee, an expert committee, and an executive board.

Users committee

This committee brings together around 20 representatives from the different wheat scientific fields and countries. They meet regularly and organize surveys in order to provide one users' feedback per year.

Expert committee (IS, scientists)

The expert committee is composed of information system specialists, PIs or heads of the platforms involved in the WheatIS platforms network, bioinformaticians, and wheat scientists. They will analyse users feedback to make recommendations for the WheatIS on orientation and priorities.

Executive board (2-3 persons)

The executive board will make decisions upon the experts' committee recommendations. They will also manage the WheatIS operational activities. They will follow the activities and report to the wheat initiative scientific board.

Conclusion: Risk analysis of the wheat information system.

The WheatIS proposal can be analyzed as follows to display its strengths and weaknesses.

Strengths

- Promote data sharing, enhance data tracking, secure data in a safe data repository
- Federate data that where originally dispersed, establish data links, standardized data.
- The wheat genome is being sequenced, generating a lot of new data that needs to be shared.
- Many recently funded large collaborative projects will provide huge amounts of data (e.g. WISP, Breedwheat, Speed, T-CAP).
- Incremental implementation has a better chance to reach the goal and answer the needs

Weaknesses

- No core funding for the WheatIS
- Managing a distributed infrastructure is complicated
- Lack of established standards to work on such a complex genome
- Late deep data integration in the project

Opportunities

- Collaboration between bioinformatics platforms at the international level that will leverage progress in data management.
- Definition of missing standards in bioinformatics
- New tools development for complex genomes
- Stimulate collaboration between scientists (public, private) by providing an infrastructure to share their results.
- Accelerate crop improvement by providing wealth of consistent data sets.

Threats

- Different funding priorities from governments that could make resources difficult to find and to coordinate.
- Different groups are involved that could have different priorities (data, accessibility). Priorities could be difficult to determine
- Lack of adoption by the community preferring to build their own system to enhance their visibility.
- With no dedicated funding sustainability, resources are more limited.

Appendix 1

Expert committee for the survey and report.

Name	***Affiliation***	***Country***
Catherine Feuillet	INRA	France
Cesar Martinez	CIMMYT	International
Dave Edwards	University of Queensland	Australia
David Marshall	James Hutton Institute	UK
Doreen Ware	Gramene	USA
Eva Huala	TAIR	USA
Hadi Quesneville	INRA	France
Hirokazu Handa	NIAS	Japan
Jizeng Jia	CAAS	China
Jorge Dubcovsky	UC Davis	USA
Keith Edwards	University of Bristol	UK
Klaus Mayer	MIPS	Germany
Mario Caccamo	TGAC	UK
Paul Kersey	EBI-EMBL	International
Peter Langridge	University of Adelaide	Australia