

## WP5 REPORT ON COLLECTIONS MANAGEMENT SYSTEM. DELIVERABLE 5.3.

### TABLE OF CONTENTS

<b>WP5 Preliminary Report on Collections Management System .....</b>	<b>1</b>
1. Introduction.....	2
2. Desired functions.....	3
3. Management system for curators .....	6
3.1. Management of collection's data using desktop applications.....	6
3.2. Management of collection's data using web based applications .....	7
3.3. Create bespoke management software using in-house resources.....	8
3.4. Use of existing open-source or free software.....	9
3.5. Use existing commercial software .....	10
3.6. Choice of databases .....	12
3.7. Backup of databases .....	14
3.8. Installation of software, versioning Information and Technology (IT) resources needs .....	15
3.9. Hosted solutions .....	15
4. Publication of data for third parties and Interoperability .....	16
4.1 StrainInfo .....	17
4.2 World Data Centre for Microorganisms .....	18
4.3 Taxonomic databases .....	18
4.4 Global Biodiversity Information Facility.....	19
4.5 Molecular and associated data resources .....	19
5. Proposal for the future Q-COLLECT infrastructure .....	21
5.1 Q-Book systems.....	22

## 1. INTRODUCTION

Collections are dealing with an increasing number of objectives and duties. Clients want more information about the biological material, improved quality services for identification, the choice of a large panel of biological material, order biological material online and rapid delivery, etc. At the same time, funding bodies are increasing their expectations requiring more scientific publications of high impact factor, increased security, tracking of the origin of the biological material, while maintaining or improving the overall quality of the collection. Of course, all these objectives should be achieved, with reduced staff and money. As if these problems are not sufficient, our (taxonomic) science is in a deep mutation phase with the introduction of new technologies that are producing large amounts of data. More and more, collection staffs have to handle increasing amounts and diversity of data quickly with limited resources.

While for decades, culture collections or museums were considered as core facilities to access type or reference material, there is a current trend to consider them as less important since data acquisition on newly sampled material is becoming cheaper and easier when using new methodologies such as next generation sequencing (for example). Therefore, biological material or specimens maintained in collections are considered by some as useless which is a major mistake since a lot of studies have been done on the latter material and a lot of publication or data are associated with them. This makes previously stored material extremely valuable for future research work.

Of course there are many different types of collections. Some are working with dead material, others with living organisms requiring a lot of attention. Some have huge amounts of specimens with little data associated while others have less strains or material but maintain large databases of associated morphological, physiological, molecular, ecological or other types of data. Some collections are barely maintained by a single person and at the same time others have up to 100 employees working on it. This means that the financial resources of collections can range from virtually nothing to several millions of euros per year which implies drastic differences in the way data are retrieved, managed, used and published. Solutions proposed to collections having nothing in electronic format, starting digitalization or having everything in databases managed by advanced software, should be quite different and adapted to their level of “digital evolution”.

This document intends to list and discuss informatics infrastructure needs for collections, their curators, associated technicians, researchers and clients or end-users. In order to find the best model or system to handle collection's operations the expected features demanded by curators and their clients are listed first and then the possible technical options that could be implemented for the curation, the publication and the use of collection's data are discussed. At the end of the document we will propose a roadmap to the different types of collections as a function of their existing digitalization status (from least to most advanced) to allow all of them to progress to the same ideal level.

To create a modern and advanced tool for the management, the analysis, the publication and the interoperability of data, there are a number of important prerequisites that need to be present:

1. Well-structured database
2. Data must be consistently and properly coded and stored
3. Well curated and maintained database
4. Database must be as complete as possible and missing data should be limited

These points form the foundation of everything discussed in this paper. Without them, the whole system would be built on a weak and unstable base.

A series of topics are presented with the advantages and disadvantages of the different possible solutions. Some of the answers are partial and slightly subjective but are needed to stimulate the development of the future Q-COLLECT information system and its outreach globally.

The document is divided into 4 major sections:

1. Management systems to assist collection curators
2. Publication of data for third parties
3. Interoperability and data analyses
4. Elements for a possible future Q-COLLECT infrastructure

The second and third sections are somehow related but deserve a separate treatment.

## 2. DESIRED FUNCTIONS

Collections data management systems (CMS) must include functionalities that will be useful to curators, technicians, researchers from the collection, clients buying the biological material and end-users of the website wishing to get data for their studies.

One should clearly distinguish features needed by curators, researchers or technicians managing or working on Collection's databases and the features needed or wanted by the Collection's clients or end-users.

Clients of a Collection are usually looking for biological material (biological material will be used further in the text to include living or dead strains, specimens, slides, or any similar reference material) that have a number of properties and want to order them quickly via an order form available from printed or, more likely now, web based catalogues. They usually want to know how much biological material will cost and when and how they will be delivered. Previously, Collection's catalogues were the only way to list all the biological material and provide additional data to clients. Nowadays such printed catalogues are abandoned and many large Collections have websites that contain the list of available biological material with some additional features. Many Collections still do not provide more data than those previously disclosed in printed catalogues. However, there is certainly a trend to increase the amount of data associated with each strain since it gives serious added value to the biological material. Most Collections allow searching for basic strain data such as strain number, species name, country of origin, substrate or equivalent collections numbers in other collections. Few collections allow clients to query their databases by multiple criteria, e.g. morphology, physiological, chemistry, molecular, ecology, geo-localization, pathogenicity, invasiveness, bibliography data or other properties.

"Researchers" interested in using Collection's data might not be clients (yet) and might use Collection's websites and associated databases to retrieve specific information, to perform correlation analyses or identify their unknown biological material against one or several reference databases. Collections that have created websites that are more than just online catalogues are more likely to attract more traffic and therefore clients than the ones just posting basic strain data without any additional tools or features recurrently attracting "Researchers". Good examples of such websites are the CBS-KNAW or MycoBank websites offering online pairwise DNA sequence alignments against

reference and curated databases. MycoBank attracts between 1200 and 2000 unique users per day by offering a number of (free) tools that allow researchers to find some solutions to their problems. Other websites such as BOLD or Genbank attract even more users by providing extremely useful functionalities.

To be helpful for the previously cited categories of users, six major lines of tools must be present and integrated in a CMS.

A non-exhaustive list of some of the desired features for the curation of Collections follows:

1. Data retrieval
  - a. Administrative data including data from collectors, depositors, geographical or ecological origin, etc
  - b. A laboratory information management system (LIMS) module to manage and track DNA sequencing projects including revival of biological material from collection stocks, DNA preparation, PCR, gels, viewing, aligning and editing DNA sequences, and depositing consensus DNA sequences into the database and online catalogue
  - c. Various biochemical data retrieval tools
  - d. Morphological characterization tools
  - e. etc
2. Data importation and exportation
  - a. Ability to import and export data as text, images, DNA trace files, microplate reader data, MS-Excel, HTML, XML, FASTA, NCBI and more
  - b. Reporting functions allow export of data in many formats including tab delimited, text, MS-word, PDF, MS-excel, HTML, FASTA, NCBI, etc.
  - c. Import, manage, analyse and export spectral data such as MALDI TOF or other systems
  - d. etc
3. Data storage and management
  - a. Advanced security and access management
  - b. Tracking of database modifications by each user
  - c. Biological material stock management
  - d. Customer information management
  - e. Orders and invoices management
  - f. Ability to create custom layouts/templates such as invoices, catalogues, sample labels
  - g. Scripting tools to automate routine tasks and extend functionalities of the software
  - h. Integration of scripts within existing menus of the software
  - i. Storage, editing and analysis of DNA and protein sequence data, including pairwise and multiple alignments, BLAST alignment of public or custom databases for identification and classification
  - j. Storage of data of many formats including texts, dates, calculations, literature references, administrative and collection data, DNA sequence trace files, electrophoresis gel photos, GPS coordinates, microplate reader data (96 or 384 wells), and photos. Data types can thus include morphological, physiological, molecular, chemical, ecological, geographic, and literature reference data
  - k. etc
4. Data analysis
  - a. Polyphasic identification and classification, to identify and classify biological material based on a custom weighted combination of DNA sequence, physiological, morphological and other
  - b. Biological material or species levels determinations

- c. Cluster analysis using various algorithms such as UPGMA, WPGMA, Single and Complete Linkage, Ward's Minimum Variance, and Neighbour Joining
  - d. Dendrogram generation
  - e. Pairwise DNA sequence alignment.
  - f. Multiple DNA sequence alignment
  - g. Generation of dynamic geographic distribution maps using Google Maps or similar tools
  - h. etc
- 5. Data publication
  - a. Direct access to published data. This means that changing data from the management software can easily and quickly be made available to the website
  - b. Easy release of new biological material and associated data
  - c. Restrict data access to Internet users/clients if needed
  - d. Easy adaption of webpages and website content. Information additions, deletions and updates should be fast and easy
  - e. Websites should be seen as a way to communicate with clients and end-users. This could be done by simple webpages/blogs, forums or news systems
  - f. Change the look and some functionalities of the website on the fly without the intervention of website developers
  - g. Allow deposit forms to be filled by depositors of biological material without having to re-type all data manually. However, they still want to control deposited data and be able to correct them if needed
  - h. Allow clients to be registered on their website and know all the information needed to contact them and send cultures with their invoices
  - i. Allow clients to easily select biological material to be ordered via a Cart system
  - j. Know pending orders, payments and data associated with any client
  - k. Allow end-users searching their databases according to the specificities of their collection
  - l. Allow third parties to take advantage of their Collection's data to increase traffic to their websites. This can be done via friendly URLs, simple or advanced web services (REST, SOAP, etc.).
  - m. etc
- 6. Data exchange/interoperability
  - a. Linking or exportation of data to other websites such as GBIF, StrainInfo, NCBI, EPPO and many more.
  - b. etc

Some of the features desirable for the "researchers" or clients of the CC and website:

Easy searching system on as many features as possible, separately or at the same time (Google like queries)

1. Advanced query system allowing to combine queries in complex ones using AND, OR and NOT operators (including brackets to group conditions)
2. Simple Cart system allowing selection of biological material to be ordered online
3. Not having to retype all personal or institutional information each time they order biological material
4. Fast and easy communication with curators or sales departments of the Collections
5. Frequently asked question (FAQ) section answering most of their questions
6. Easy copy-pasting of data
7. Easy exportation of selected data, manually or via software (web services)

8. Pairwise DNA or protein sequences alignments against reference databases
9. Polyphasic identifications and/or classifications against reference databases
10. MLST (or similar methods) allowing identifications or typing of biological material
11. Forum to discuss questions related to the community of users
12. Online support
13. etc

There are many more features that could be listed above and this list is certainly non-exhaustive and will grow over the years.

### 3. MANAGEMENT SYSTEM FOR CURATORS

To create an efficient and advanced data storage, analysis and publication system for Collections, a number of different technological options are possible but some present more advantages than others. These are discussed here. This section is critically important if one wants to create a data system that will be used by third parties in an efficient way. Before thinking of creating high level applications and interoperability scenarios, each participating culture collection must have a well-structured data management system. Without the right foundation (structure, data, software tools and IT infrastructure), it is impossible to create basic or advanced functionalities that will position Collections adequately in a modern scientific and interoperable landscape.

#### 3.1. MANAGEMENT OF COLLECTION'S DATA USING DESKTOP APPLICATIONS

Desktop applications (DA) constitute the majority of software that is available at the moment. Software such as Word, Excel, and Access from Microsoft are typical DA. Many of the collections are currently using DA to manage their collections with Excel being the tool of choice for the smaller Collections, as it is easy to use and understand. It contains a lot of functionalities that might be useful for a large number of operations. Collections that need more advanced systems might use Access or FileMaker Pro. Unlike Excel, the latter systems are multi-users and relatively easy to use without real programming skills.

Table 1. Advantages and disadvantages of desktop applications

Advantages	Disadvantages
Rich software interface	Installation can be problematic (different Operating System (OS) versions, missing Dynamic-link library, etc.)
Easy to use	DA are usually made for one OS (Windows, Mac or Linux) but won't work with others
Fast response to user's commands	When installed on different computers, updates and upgrades of the software must be re-installed everywhere making bug fixing or new version less easy to fix or install
Memory demanding or interface rich operations can easily be performed (to the technical limits of the OS, computer, etc., of course)	DA are usually not accessible from a remote computer or device
Relatively easy to develop (for basic functionalities at least)	For software working with limited installation options (fixed number of licenses), DA might become expensive and/or difficult to update/upgrade

Interactions with other software can be easy to establish. Pipelines can be created and import-export functionalities easy to implement or to use	Can be heavy to manage for IT departments
Data access security can easily be ensured	

DAs remain the dominant systems to access and manage collection's data. They are easy to use and fast but installations and software maintenance can be challenging, especially in collections with multiple curators or users (technicians, researchers, etc.) using different OS. Also, in such a multi-user environment, connections to the central database must be relatively fast in order to avoid slow responses or disconnections and consequent data losses.

All of the above mentioned disadvantages can be alleviated by using application servers and remote desktop access (RDA) software such as RDP from Microsoft or Citrix XenApp or XenDesktop that are even more efficient in terms of memory usage, speed and display quality (other RDA systems are also available). Currently, using Citrix to publish DA is certainly the best possible combination and allows access to a rich and fast interface on any OS with any version and on any device (Desktop, laptop, tablet, smart phone, etc.) from anywhere. Installation is central and updates or upgrades are easily published.

### 3.2. MANAGEMENT OF COLLECTION'S DATA USING WEB BASED APPLICATIONS

Web based applications (WA) constitute a good alternative to DA. They are typically accessible using a browser that can be found on any device using any OS.

Table 2. Advantages and disadvantages of Web based applications

Advantages	Disadvantages
Accessibility to databases from anywhere	Development costs can be higher
Accessibility to databases from multiple platforms	Developments can be significantly more complex to support all browsers and their versions
Possibly easy to use for basic editing of data	Some functionalities are more difficult or impossible to program even if this is becoming less and less the case
Maintenance is easy for IT departments since the software is centrally installed and maintained	Rich interfaces or memory demanding operation might be impossible
No need for installation on curator's, researcher's or technician's devices (Desktop, laptop, tablet, smart phone, etc.) since access is ensured via browsers	Interface can be much slower than DA
The same software might be used for the management and the publication of data	Interactions with other software might be more difficult or impossible
	Maintenance of software might be more intensive to allow new versions of browsers to still function properly
	Security issues are more complex to handle with WA than with DA since the application is potentially accessible from any device by anyone
	Stable Internet connections are needed

While the advantages listed above seem attractive, currently, WA remain too slow and limited in their functionalities and capacities to handle some specific data. Technological advances (.NET, Java, Silverlight, HTML 5, etc.) might

resolve some of the issues mentioned above. As an example, Microsoft office is now partly available in a web based form and many desktop features are also available in the web based version.

Some culture collections have moved from DA to WA but the majority of them are still using DA for the management/curation of their databases. So for the management and curator's operations, it seems that DA remains the best choice for the moment but this might change in the future.

### 3.3. CREATE BESPOKE MANAGEMENT SOFTWARE USING IN-HOUSE RESOURCES

A number of large Collections have developed their own systems to manage their data. This is certainly a possible solution when good and stable programming skills are easily accessible.

Table 3. Advantages and disadvantages of in-house software developments

Advantages	Disadvantages
Tailor made application fitting perfectly with the needs of the curators (at design time at least)	Curators or researchers are rarely good software designers or programmers making the resulting solution uneasy to use, maintain and further develop
Fast response to implement new features and bug solving	Real developers are rarely available in Collections because they are expensive
This solution can be quite cheap if the software remains simple	Good developers tend to leave the Collections to find better paid positions leaving the software unmaintained and hardly usable by newly recruited developers
	This option can be extremely expensive when the wanted functionalities are complex and large
	Most in-house solutions are not (easily at least) scalable (add/modify/remove more tables, fields, operations, etc.) and redesign or complete; rewriting of software is often needed. This leads to interfacial instability for the users which is a key issue
	Developments take a long time before being usable and stable especially for single or small developer's teams
	Many software were abandoned after a few months because they were too slow, difficult to use, user-unfriendly, buggy or unstable. This is a common situation in a Collections

While for a very small Collections with one or two users (curator/researcher, technician), this can be seen as a viable solution provided that the system to be developed remains simple, it is certainly not an advisable solution for most Collections particularly when serious teams of developers are lacking. Note that human IT resources have to be distinguished from developers. They have quite different skills, though overlapping sometimes a little. It is a common mistake to confuse IT people with developers and this often leads to disappointment when IT staff are forced to program data management or (even worse) analysis software.

Therefore, we strongly suggest Collections (small or big) to develop their own CMS but rather to use free or commercial third parties solutions.



### 3.4. USE OF EXISTING OPEN-SOURCE OR FREE SOFTWARE

Using open-source or free software is really common among Collections due to the lack of financial resources to buy commercial solutions. Many tools have been developed to manage, analyse and publish data. Some are easy to use and propose very interesting functionalities. A typical example is the BLAST software family that allows aligning sequences very efficiently which is one of the many operations that curators are doing on a regular basis. Many other excellent open-source or free software can be listed that can perform basic or even advanced functionalities requested by curators in their daily operations. They include BLASTN, BLASTP, BLASTX, Geneious (entry version free), Mantis, Mega, RasMol, Scratchpads, SeqView, Serial Cloner, Specify, World Data Centre for Microorganisms (WDCM) workbench, etc. (a far from exhaustive list)

While some of the solutions are extremely efficient in their field, there is no open-source or free solution that can handle all the operations that are needed by curators. However, some solutions are quite interesting such as ScratchPads (SP) and the newly developed WDCM workbench (WB) created by the Chinese Academy of Sciences (CAS).

SP were created by a researcher of the Natural History Museum in London; their website states “Scratchpads are an online virtual research environment for biodiversity, allowing anyone to share their data and create their own research networks. Sites are hosted at the Natural History Museum London, and offered freely to any scientist that completes an online registration form. Sites can focus on specific taxonomic groups, or the biodiversity of a biogeographic region, or indeed any aspect of natural history. Scratchpads are also suitable for societies or for managing and presenting projects. Key features of Scratchpads (see also Scratchpads feature list) include: tools to manage biological classifications, bibliography management, media (images, video and audio), rich taxon pages (with structured descriptions, specimen records, and distribution data), and character matrices. Scratchpads support various ways of communicating with site members and visitors such as blogs, forums, newsletters and a commenting system.”

SP are more oriented towards the management of museum data and are therefore lacking a number of features that are absolutely needed for other types of Collections, such as stock management, orders management, and other advanced tools that are used on a daily basis by curators. For example, tools that support electrophoresis, microplate management, MALDI-tof, DNA and Protein sequences management tools and many more.

Since SP are free, support is quite limited and additional tailor-made developments are not possible from the developers.

The WB is an interesting initiative from CAS and intends to propose a ready to use system for the management and the publication of Collection's data. The system is hosted at the CAS and a few small to medium collections are using it although the system is still in its early stages of development. The system is fixed by nature (not dynamic) which means that fields and tables cannot be added by the curators of the collections in order to correspond to their own needs. This is certainly a major issue since each collection is specific and hosts different types of organisms/material and therefore data requiring significant differences in fields and tables. WB can however be an interesting solution for a small collection with limited resources particularly if the data sets defined in the OECD Best Practice Guidelines for BRCs (OECD 2007) or the CABRI guidelines ([www.cabri.org](http://www.cabri.org)) are used i.e. the Minimum Data Set (MDS), Recommended Data Set (RDS) or Full Data Set (FDS) relevant for each group of microorganism. This solution is not working for non-microorganisms making it not practicable for Q-COLLECT.

Open-source software can be of interest for collections having serious teams of developers but as a general rule, using the code of third parties is often a real challenge, especially for large software. Even experienced developers can struggle to understand the code written by others even if the code is well documented which is not always the case. The major advantage of Open-source software remains the ability to add missing functionalities to already existing and almost perfect software. Unfortunately (to our knowledge) there is no such complete (open-source) solution that could be used for the management of Collections.

Table 4. Advantages and disadvantages of open-source software solutions

Advantages	Disadvantages
Ready to use	Some solutions can be free and open source but could become expensive if major coding changes are needed
Known solution with known limits and advantages from the beginning	All open-source solutions are not of equal quality in terms of code
Code is accessible to anyone and can be changed if needed	Code can be hard to extremely hard to understand and maintain even by good developers
Free of charge	No professional support. Sometimes no support at all for poorly used software
Take advantage of solutions developed by others	
Community based support	
Non-dependency to software company	

Free software (non-open source) can be of interest of course but here again there is no free solution that would fit all needs. Using a large number of different software complementing each other can be a solution but this option is often less efficient than a completely integrated system fulfilling all or almost all the needs of curators. Some pipeline software (integrating different software) such as Taverna (there are many others as well) can be used to better integrate several individual software by joining the inputs and outputs. This is certainly a solution for some scenarios but not a viable solution for a complete Collection management system.

### 3.5. USE EXISTING COMMERCIAL SOFTWARE

There are a few commercial software options that could be used to manage all the operations associated within a Collection. Again, a non-exhaustive list is presented: BioloMICS, Bionumerics, FileMaker Pro, Geneious, KE Emu, LabCollector LIMS, MS-Access, MuseumPlus, Oracle, Etc.

There are many commercial software options that were specifically created for the management of museums operations and two of them are cited above (KE EMu and MuseumPlus) since they seem to be among the most popular ones. Those solutions are used by major players in the museum arena but Collections needs are slightly different and certainly more extensive since most Collections are dealing with data such as morphology, physiology, chemistry, ecology, molecular and many more that are usually not handled by museum targeted software.

Other software such as FileMaker Pro, MS-Access, PostgreSQL or Oracle are based on databases and can be completely or partly programmed to fit the needs of curators of Collections. Here again, programmers are needed to create a complete and functional package and once again, many of the needed functionalities are not present by default in those software.

Laboratory Information Management Software (LIMS) belong to another family of solutions that can in some cases provide an accurate but partial and often very expensive solution. LIMS are made to track samples, experiments, etc. and provide advanced reporting solutions. Some collections such as the BCCM have followed this route to handle some of their operations and are implementing a LIMS solution provided by Siemens. Such systems are extremely expensive (buying and maintenance costs) and require a lot of investment in terms of adaptations to the needs and specificities of the different Collections. LIMS alone do not provide all the functionalities needed by Collections.

Some software like BioloMICS, Bionumerics or Geneious propose advanced solutions that can fit with some or all (depending on the complexity) the management needs of a Collection.

Geneious is a relatively new player and is a “DNA, RNA and protein sequence alignment, assembly and analysis software platform, integrating bioinformatics and molecular biology tools into a simple interface” (from website). It offers a wide-ranging functionality:

- Access essential molecular biology analysis tools and plugins, and search public and private databases, all from one location
- Organized Data, step into the future with simple drag and drop import of a vast range of formats. Arrange and browse your data library how you like
- Superior Visualization, switch to a clear and bold graphical interface. Eliminate the need for command-line operations and stop battling with poorly designed software.

A LIMS module is also available for the management of 1<sup>st</sup> generation sequences. Another one can handle NGS data. This system allows importing data from a number of databases but is not genuinely accessing database records directly. In fact, Geneious is a great tool integrating a number of analysis modules but cannot be considered as data management software that can be used for all basic Collection’s operations.

Bionumerics software features a very large number of analytical modules capable of analysing a large number of data types. From this point of view it is probably one of the most complete since morphological, chemical, physiological, electrophoresis, spectroscopic or molecular data can be used to identify or classify biological material or species records. Bionumerics is used by a very large number of laboratories, including Collections. This software is mainly used for its analytical features. However it can import and handle data from many database sources. It also offers a scripting tool using its own language. Bionumerics developers can also write scripts in Python for their customers at an hourly rate. This is a major and important feature allowing the customization of the software but the language used is non-conventional and cannot offer all the advantages of modern programming languages such as Visual Basic, C#, C++ or Java (for example). Bionumerics imports records from existing databases in order to analyse them but no queries can be performed which makes it impossible to use for most of the common operations of a Collection.

BioloMICS was first created almost 25 years ago to manage yeast collections and perform batch morphological and physiological identifications. This software went through a large number of iterations and the current version is 10. “BioloMICS is the most complete software solution for storage, management, analysis and publication of biological data and is of choice for any research or industrial laboratories, museums, culture collections and many more” (from website). Any data type can be stored and handled in BioloMICS, from morphology, physiology, biochemistry, chemistry, chromatography, electrophoresis, molecular to bibliography, taxonomy, geography, ecology or administrative. The data structure is fully flexible. One can very easily create tables and fields (24 different field types can be used to manage all possible types of data) of interest on the fly and handling data of any kind. The system keeps track of all the changes ever made in the database. The system currently uses MySQL for the underlying

database but a new version under preparation will allow using MSSQL, PostgreSQL or MongoDB for very large datasets. Data cannot only be managed but also analysed in a similar way to Genieous and Bionumerics. It offers a large number of tools to analyse morphological, physiological and sequence data. Polyphasic or multi-locus identifications are possible as well as clustering tools that can produce hierarchical trees or three dimensional structures. The BioloMICS software provides LIMS for the complete management and analysis of 1<sup>st</sup> generation sequencing data. The software allows writing scripts but unlike Bionumerics it uses .Net technologies like Visual Basic or C#. Scripts can be integrated in the existing interface allowing the extension of the functionalities of software to fit with the particular needs of the end-users. Recently a debugger and a form designer have also been integrated that make this software the most complete of its kind. Support is also efficient and developers can write tailor-made programs at an hourly rate. This software has been created for collections and is now sold as a commercial product to a number of Collections world-wide such as CBS-KNAW, CABI, Pasteur Institute, CDC, University of California, almost all Australian microbial collections and many more. It is certainly the most complete software for Collections for the moment since it also includes a web publication interface that is used by a number of large international initiatives such as MycoBank, the European Barcoding Database Mirror or Q-bank, which is major advantage in the framework of the Q-COLLECT project.

Commercial software are usually more expensive by definition since they are not free. On the other hand they are ready to use and there is no lag phase between the buying stage and the moment where curators can have a functional system. Collections that have developed their own projects often abandon them due to the lack of ability to achieve their primary goals or the inability of their software to cope with new types of data or operations. It is finally often cheaper to buy or to rent commercial software and pay for the updates rather than supporting expensive developers within a Collection. Curators are neither software developers nor software designers and cannot properly manage or guide teams of developers. Therefore, software developed within a Collection can be badly designed and take a long time before being usable.

Table 5. Advantages and disadvantages of commercial software solutions

Advantages	Disadvantages
Ready to use	Some solutions can be expensive and sometimes extremely expensive
Known solution with known limits and advantages from the beginning	All commercial solutions are not of equal quality and not all are suitable for a Collection
Software are usually well-written and maintained by professional developers	When access to the databases is not possible via scripting or via database direct access, specific developments can be impossible and this is a major issue for possible future extensions and needs
Can be much cheaper in the long term than paying software developers internally	Dependency on the software and the company producing and maintaining it
Take advantage of solutions developed by others	
Professional support	

### 3.6. CHOICE OF DATABASES

Different types of databases or supporting tools are used, some of these are listed below in increasing order of complexity or capacity: Catalogues on paper (not a database sensu stricto but still used); Word processing software (not a database but used by a number of collections); Excel (not really a database but still used by a large number of

Collections; inexpensive); MS-Access (basic relational database; inexpensive); FileMaker Pro (basic relational database; not free and associated with the management software); MySQL (simple relational database; free); PostgreSQL (relational database; free); MSSQL (relational database; not free, not cheap); Oracle (relational database; not free and expensive); MongoDB (document oriented database; free); Vertica (grid-based column oriented database; expensive); Etc.

While a few Collections are still managed by paper cards systems or by paper like catalogues (see report of WP2 of Q-Collect), some are using word processors to keep track of the information associated with their biological material. Such systems are of course outdated and should certainly be replaced by more efficient tools that can be used to efficiently manage Collection's data and publish them on dedicated websites. The number of Collections operating with such outdated systems is certainly not negligible as seen by surveys made by other WP of the current Q-COLLECT project.

While Excel cannot be considered as a real database it certainly delivers a number of advantages and interesting features to a very small Collection. It is certainly not the system of choice to manage medium to large Collections with more than one curator/technician.

MS-Access can be considered as a relational multi-user database. "Microsoft Access stores data in its own format based on the Access Jet Database Engine. It can also import or link directly to data stored in other applications and databases. Software developers and data architects can use Microsoft Access to develop application software, and "power users" can use it to build software applications. Like other Office applications, Access is supported by Visual Basic for Applications, an object-oriented programming language that can reference a variety of objects including DAO (Data Access Objects), ActiveX Data Objects, and many other ActiveX components." (source Wikipedia). It offers a number of advantages over Excel but the system is moderately supporting simultaneous updates and therefore, it should not be recommended for medium to large Collections.

"FileMaker Pro is a cross-platform relational database application from FileMaker Inc., formerly Claris, a subsidiary of Apple Inc. It integrates a database engine with a GUI-based interface, allowing users to modify the database by dragging new elements into layouts, screens, or forms. Current versions are: FileMaker Pro 12, FileMaker Pro Advanced 12, FileMaker Server 12, FileMaker Server Advanced 12, and FileMaker Go 12 for iPhone and iPad. FileMaker evolved from a DOS application, but was then developed primarily for the Apple Macintosh. Since 1992 it has been available for Microsoft Windows as well as Mac OS/OS X, and can be used in a cross-platform environment. FileMaker server briefly ran on Linux, but Linux support was abandoned with FileMaker 7, and the server currently runs only on Windows or OS X servers. It is available in desktop, server, iOS and web-delivery configurations. FileMaker, since version 9, includes the ability to connect to a number of SQL databases without resorting to using SQL, including MySQL, SQL Server, and Oracle. This requires installation of the SQL database ODBC driver to connect to a SQL database. SQL databases can be used as data sources in FileMaker's relationship graph, thus allowing the developer to create new layouts based on the SQL database; create, edit, and delete SQL records via FileMaker layouts and functions; and reference SQL fields in FileMaker calculations and script steps. It is a cross platform relational database application." (source Wikipedia). FileMaker Pro has been used by a number of Collections thanks to its ease of use and flexibility.

MySQL, PostgreSQL, MSSQL and Oracle belong to the same family of relational databases that are used by most of the medium to large size Collections. Such databases offer a wide range of possibilities and present advantages and disadvantages. MySQL is certainly one of the most used since it is free, easy to use and fast. However, MySQL does not offer all the tuning tools and programming interfaces that PostgreSQL, MSSQL and Oracle can offer. All these

databases can handle most of the datasets that small, medium to large Collections have to deal with. They are probably the solution to 99% of the datasets management issues at the moment.

“MongoDB (from "humongous") is an open source document-oriented database system developed and supported by 10gen. It is part of the NoSQL family of database systems. Instead of storing data in tables as is done in a "classical" relational database, MongoDB stores structured data as JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster.” (source Wikipedia). MongoDB could be a good solution for very large and distributed datasets. Very few software mentioned above are able to use or connect to such a database. This will probably change in the near future due to the need to handle very large datasets produced by high-throughput systems (NGS for example).

There are many other database types. One of them is Vertica that is a grid-based, column-oriented database. “Vertica Analytic Database is designed to manage large, fast-growing volumes of data and provide very fast query performance when used for data warehouses and other query-intensive applications.” (source Wikipedia). Scenarios where such databases could be used remain extremely marginal in the world of Collections but the so-called “Tsunami of data” problem might push some Collections to adopt such extreme technical solutions.

Data standards are not discussed here since they should be considered as more or less independent from the databases in which they are stored. The way data are stored in databases is usually depending on the software managing and using them. There is certainly no strong reason to enforce some specific formats at this stage (see the Interoperability section for more on standards) but agreement on data exchange is certainly an important factor that must be accounted.

### 3.7. BACKUP OF DATABASES

The system chosen must include a systematic automated backup procedure because humans tend to forget to do them or do so at irregular intervals. Most database systems integrate automated backup procedures. Backups should not be stored on the same computer or server as the running database. Ideally, some backups should be stored on one or several remote computers or servers in order to prevent problems related to computer failure, power problems, fires, etc.

A good practice is to backup databases once or twice a day and keep all the versions for one week; keeping copies on a weekly and monthly basis.

Some databases can be stored on several database servers in order to propose a highly available system (redundancy). Other databases can also do *Sharding* which is the process of storing data records across multiple servers, some records being stored on some machines while others will be stored on others. The management of such systems is usually done by the database engine. MongoDB (and several others) makes use of such very interesting options.

Depending on the chosen system/software, some files might not be included in the database as blobs but are stored in the file systems. In such a case a database backup is not sufficient and one must also backup all the files associated with the records in the database.

### 3.8. INSTALLATION OF SOFTWARE, VERSIONING INFORMATION AND TECHNOLOGY (IT) RESOURCES NEEDS

Different Collections are working with different software systems and different operation systems. However, the vast majority of computers are still working under Microsoft Windows (XP, 7, 8 or 10 and equivalent versions for the servers). Some are using Mac OS while a very limited number of Collections might use Linux. The following statistics obtained from Netmarketshare website (sources <http://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustommd=0>) allow to objectivate the assumption above: 90.7% of the computers and therefore software are running under Windows, 7.7% with Apple IOS and 1.6% with Linux. Therefore, the system recommended for the management of Collections should be able to work properly (at least) under a Microsoft Windows OS. Creating or selecting desktop software that can work on all existing OS is an unnecessary challenge or burden. Therefore it is strongly recommended that the chosen software is capable of running on Microsoft Windows.

However, if the software can be installed on an application server and served to the end-users using Citrix XenApp or XenDesktop, RDP or any similar system, then the operating system of the end-user is not a limiting factor anymore. In such a case, an Apple based OS software could be used under Microsoft OS and vice-versa. In such a scenario any device can be used by the end-user, including thin-clients. This being said, creating an application server requires hardware resources as well as IT skills and support that are above the normal level of IT support. This cannot easily be achieved for small to medium size collections.

When software is installed as a client server solution (Software on the PC of the end-user and the database on a central server), updates and upgrades of the software can be challenging and may require quite some time for IT personnel. When updates are frequent it becomes important to choose software that can be updated or upgraded automatically. Most software can now do it but this should certainly be a requirement in a client server solution.

In the application server scenario mentioned above, this is less of an issue since the software is installed centrally and update once on the central/application server. For medium to large Collections, this option is certainly the best one.

While for a small Collection, dedicated IT staff might not be needed, for medium to large Collections, it is important to be able to rely on skilled and effective IT staff. They usually take care of backups, software installations and maintenance. In the case where many servers need to be used and maintained a number of specialized software will have to be acquired in order to monitor and manage the whole system. Software such as VMWare, Citrix XenApp/XenDesktop, HyperV and many others that are expensive tools must be used to have a professional system with a high level of availability and security. Such a system is expensive to establish and to maintain. Most often small Collections are hosted in research institutions or Universities that have their own IT support.

### 3.9. HOSTED SOLUTIONS

When IT staff and hardware resources are lacking or when financial resources are limited, hosted solutions are certainly interesting and should be favoured. The hosting company usually takes care of everything for their end-users including: installation of CMS (possibly of the publication software as well); updates and upgrades of the software; installation of database(s) and file system needed to store associated documents; Backups of databases; make applications available via an application server for desktop applications; make a website available to third parties/publication of Collection's data; high availability of the system; hosting is almost always done from

professional data centres with high security standards (redundant power supply, protection against fire and thefts, firewalls, etc.).

Table 6. Advantages and disadvantages of hosted solutions

Advantages	Disadvantages
Easy to use	Require recurrent payments (monthly or annually) which means that these costs must be part of the annual budget of the CC
End-users can directly have access to a complete and efficient system with lag period	Access to a database engine might not be possible (only backups of databases are provided from time to time)
No need to buy hardware (server, SAN, firewalls, etc.)	Dependency on the hosting company
No need to buy and maintain expensive and sophisticated software for the management and the monitoring of the system (VMWare vSphere, for example)	Need Internet connection to work
No need to hire IT staff	Extremely slow or erratic Internet connections might be unable to use such a system
Continuous monitoring and support	
Given the number of services provided, hosted solutions are often much cheaper than running a complete infrastructure in house	
Access to database and software from anywhere at any time on any device	
Management of CC software and associated database can directly be connected to the website used for publication of CC data	

Few companies offer complete solutions but three can be mentioned that not only offer the software that can be used to manage a Collection but also publish Collection's data and propose a hosted system:

- ScratchPads (no desktop application to manage data but web based), no real support since it's a free service (see discussions above about the limits of this system). Hosted at London Natural History Museum.
- BioloMICS. Hosted in professional data centre. Desktop and web portal are both included. Remote access to management of the Collection's application via Citrix XenApp/XenDesktop.
- WDCM Collection's management system of the Chinese Academy of Sciences in Beijing

#### 4. PUBLICATION OF DATA FOR THIRD PARTIES AND INTEROPERABILITY

Common data management standards including adopting common ontologies are essential for interoperability between collections and outside to other types of data. The Collections community has standards for data management; the EMbaRC and GBRCN project consortia partners, the predecessors of MIRRI, decided that the CABRI guidelines could be amended and adopted by MRCs. However, in Q-COLLECT the focus is on what the user needs are and how this impacts on the stored data and thus on the ways of presenting them. Lists of fields and types of fields have been addressed, the OECD best practice guidelines for Biological Resource Centres (BRC) published in



June 2007 brings together previous work and makes appropriate recommendations (<http://www.oecd.org/health/biotech/oecdbestpracticeguidelinesforbiologicalresourcecentres.htm>). Controlled vocabularies/ontologies need to be addressed in some way.

Q-COLLECT needs a strategy on what data are really needed to help facilitate the uptake and use of quarantine data in research and development. Analyses made in WP2 on what data are out there showed that many Collections have very little data and often no catalogues or any similar listing. It is beyond the scope of the Q-COLLECT project to gather all data from existing Collections and to create a fully functional system to make them available to all. However, it the goal of the current WP to prepare the work of potential future projects to achieve this highly needed task of mobilizing data from smaller or less digitalized collections and making them available to all.

In order to link Collections data to other systems it is imperative to follow the necessary standards to allow third parties to use our data with confidence. In order to do so, software systems used by Collections need to be able to easily export or expose data and ideally automatically using a number of formats that are usually XML based and that should probably be independent from the format of the original database where data are maintained. There are many initiatives trying to establish biological data standards as well as standards that are used by biologists such as geographic, climatic or ecological data, for example. It seems that reinventing such standards is certainly not a good idea since our community does not have the ability or capacity to contribute to this. What we should certainly do is to identify a number of standards that are relevant to the type of data that Collections are likely to use and produce and ensure that the software used by Collections are able to utilise such standards. A number of websites, documents or working groups are certainly of major interest with respect to data standards:

1. BioSharing (<http://biosharing.org/>)
2. Biodiversity Information Standards (TDWG; <http://www.tdwg.org/>)
3. Genomic Standards Consortium (GSC; [http://en.wikipedia.org/wiki/Genomic\\_Standards\\_Consortium](http://en.wikipedia.org/wiki/Genomic_Standards_Consortium))
4. More are available

#### 4.1 STRAININFO

The StrainInfo (SI) portal has been gathering data at the strain level between participating Collections for several years. Originally SI was screening websites of collections but this system seemed to be inefficient and their initiators decided to create the Microbiological Common Language (MCL) which is an XML based format allowing culture collection's microbial data to be exchanged between Collections and SI. "In short, MCL defines terms which can be used to reference and describe microorganisms. It is designed to form a simple and generic framework leveraging the electronic exchange of information about microorganisms. MCL is loosely coupled from its actual representation technologies and is currently used to structure XML and RDF files" (from <http://www.straininfo.net/projects/mcl>).

SI is a useful portal since data from many microbial culture collections are centralized and compared and discrepancies between identical biological material present in different Collections are highlighted for curation purposes. SI also associates molecular and bibliographic data from NCBI and PubMed to basic strain data. SI also provides links to websites and databases where the biological material are originated from.

SI is quite comprehensive and software managing Collection's data should have the ability to export data in MCL format that can be used by SI and therefore ensure better visibility of participating Collections. Currently, MCL does not include all data held by microbial resource collections but it could easily be extended to cover all data elements

Q-COLLECT would need. This being said, SI is strictly limited to microbial data and strains present in culture collections, is not meant to cope with the broader scope of “objects” that Q-COLLECT Collections are dealing with.

## 4.2 WORLD DATA CENTRE FOR MICROORGANISMS

The World Data Centre for Microorganisms (WDCM) is based in Beijing, China and managed by the Chinese Academy of Sciences. WDCM maintains a catalogue of the largest Collections in the world. It was created “to enable broader and easier access to the reference biological material listed by the ISO TC 34 SC 9 Joint Working Group 5 and by the Working Party on Culture Media of the International Committee on Food Microbiology and Hygiene (ICFMH-WPCM) in their publication Handbook of Culture Media for Food and Water Microbiology. It fulfils a need expressed by these bodies for a unique system of identifiers for biological material recommended for use in quality assurance.” (from <http://refs.wdcm.org/home.htm>). WDCM is requesting data to be submitted using a tab delimited format. WDCM is proposing services that are very similar to the ones proposed by SI. Therefore the same concluding remarks apply.

## 4.3 TAXONOMIC DATABASES

One of the most fundamental problems of managing a Collection organisms is keeping pace with the taxonomy and resultant name changes being introduced for species. This is highlighted when databases are brought together; a specific case in point being the tremendous amount of time taken up during the integration of MINE – Microbial Information Network Europe data. This was again highlighted during the CABRI – Common Access to Biological Resources and Information project [www.cabri.org](http://www.cabri.org). The problem is still encountered by databases such as the WDCM when it lists the numbers of species (names) held by its registered collections and demonstrated by species lists and strain number linkages shown when Straininfo.net ([www.straininfo.net](http://www.straininfo.net)) is searched. There are very few tools that can cope with this centrally and to get every name right for the 2 million plus biological material in the WDCM database would be a tremendous task; several attempts have been made to do this over the years.

A number of taxonomic or nomenclatural databases are available to link Collections biological material data to currently recognized scientific names. For Fungi, MycoBank (MB; <http://www.mycobank.org>; nomenclature and taxonomy) and Index Fungorum (IF; <http://www.indexfungorum.org>; nomenclature) are the main players. For bacteria, DSMZ culture collection (<http://www.dsmz.de/>; nomenclature and taxonomy) publishes monthly updates of bacterial nomenclature and taxonomy. The latter is not searchable online but can be downloaded. Another interesting website is certainly the List of Prokaryote Names with Standing in Nomenclature available at <http://www.bacterio.net/>. The system is rich in terms of data but has serious limitations in terms of interoperability since data are not stored in a database but in html pages and there are no real web services allowing to easily link and retrieve data. Therefore, for the Q-COLLECT project, a Bacterial names search engine with associated web services has been created, working exactly like MycoBank (<http://www.mycobank.org/bacteria>).

The Catalogue of Life (CoL; <http://www.catalogueoflife.org>) initiative is another solution to get access to taxonomic information that is not just specialized for Fungi or Bacteria but integrates higher organisms as well.

The Encyclopedia of Life project (EOL; <http://eol.org>) is yet another database with a nice website offering species descriptions and associated metadata on the many life-forms on Earth - of animals, plants, fungi, protists and bacteria. Like CoL, EOL is an aggregator of data obtained from other databases such as MB or IF, for example.

EPPO Global Database is maintained by the Secretariat of the European and Mediterranean Plant Protection Organization (EPPO). Although the main focus is to provide information on regulated pests the system includes organisms important in agriculture and crop protection: crops, pests (including pathogens and weeds), natural enemies, and organisms used in ecotoxicological studies. It includes the core data files of the Bayer codes which were previously available as a set of books or data files. This coding system is now managed by EPPO and Bayer codes can now be called EPPO codes. Although this system was built for agronomic purposes for each organism, it provides taxonomic elements (main steps of a taxonomic tree, preferred scientific names and synonyms). In addition, this coding system contains many common names in different languages. At present, the database covers more than 68 000 species (plants, animals and microorganisms)

For the management of names, Collections should not maintain their own nomenclature and taxonomic databases since this task is far too complex and would require important dedicated resources that are, most of the time, not available. MycoBank is an example that is delivering a number of web services that can be used to link Collection's biological material to a central and curated system (<http://www.mycobank.org>) that should be followed.

#### 4.4 GLOBAL BIODIVERSITY INFORMATION FACILITY

The Global Biodiversity Information Facility (GBIF) is an integrator system that centralizes data from a diversity of resources including major Collections. Data from different museums, collections, nomenclators and others are combined into their system and linked on the basis of their geographical or ecological origins. Data can be queried and links to the original websites are provided to get more information on the interesting records.

GBIF aggregates more than 400 million data records and is therefore a serious source of information for people working with biodiversity related matters.

Exports to GBIF are usually done using a Darwin Core archive format (DwC). The information system that will be chosen by Collections should therefore offer the ability to export to DwC.

#### 4.5 MOLECULAR AND ASSOCIATED DATA RESOURCES

Most molecular data produced by researchers worldwide are deposited in one of the three International Nucleotide Sequence Database Collaboration (INSDC):

- NCBI-GenBank in the USA ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov))
- EMBL in the UK ([www.ebi.ac.uk/ena](http://www.ebi.ac.uk/ena))
- DDBJ in Japan ([www.ddbj.nig.ac.jp](http://www.ddbj.nig.ac.jp))

Most sequences (DNA or Proteins and associated metadata) are regularly synchronized between the three databases and the main part, that is available from the first database, is available from the others as well.

INSDC databases are major sources of genomic and metagenomic information and links to and from them are of key importance to any Collections. Exportation to and importations from INSDC database tools must be available in the software system managing Collection's data.

The Barcoding of Life Database (BOLD; [www.barcodinglife.com](http://www.barcodinglife.com)) is a major international initiative that was started a few years ago and that has gained a lot of popularity in recent years. BOLD is focused on DNA barcoding and most of the available data are related to higher organisms. Very few microbes are represented in their databases but the intention is certainly to include more of them. This is likely to evolve in the future since a number of fungal institutions such as CBS-KNAW are dedicated to produce large numbers of fungal ITS sequences in the near future and the latter will be submitted to BOLD and GenBank.

CBS-KNAW and Naturalis have launched the European mirror of the BOLD system but unlike the latter it includes much more fungal ITS sequences that could be used for identification. The BioloMICS software used by CBS-KNAW for this mirror and for MycoBank allows Collections using this software to create a portal that can be accessed remotely and used to perform pairwise sequence alignments against Collections that would like to share their DNA sequence databases. This system attracts visitors to the Collections and can potentially increase visibility and initiate business opportunities.

ELIXIR is a major EU funded project that “unites Europe’s leading life science organizations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information. ELIXIR will provide the facilities necessary for life science researchers - from bench biologists to cheminformaticians - to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built. The purpose of ELIXIR is to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society. The collection, curation, storage, archiving, integration and deployment of biomolecular data is an immense challenge that cannot be handled by a single organization or by one country alone, but requires international coordination. ELIXIR will provide the facilities necessary for Europe’s life science researchers to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built. In order to achieve its mission, ELIXIR will construct, operate and enhance a distributed research infrastructure in accordance with the requirements of the scientific community and under the direction of the ELIXIR Board. The ELIXIR Hub will be connected to ELIXIR Nodes to provide infrastructure for data, compute, tools and standards and training as well as support for the ESFRI biological and medical science infrastructures.” (from <http://www.elixir-europe.org/>). As far as we are aware, the ELIXIR system will be a distributed system of resources that will be usable for specific purposes. Specialized consortia will produce software or combine new or existing ones to allow answering specific questions such as for example: screening organisms for solutions or products for the market; e.g. ways to accelerate the discovery of new antimicrobials where one may have uncharacterised organisms or even microbial diversity in the soil where nobody has any idea of their potential.

Q-bank (<http://www.q-bank.eu>) is another international initiative related to the barcoding of quarantine related organisms that links, among other types of organisms, microbial data to DNA barcodes and quarantine related data. We believe that projects resulting from the current Q-COLLECT initiative should base their software and databasing infrastructure on the Q-bank system since it offers most of the functions wanted by Q-COLLECT curators and end-users.

## 5. PROPOSAL FOR THE FUTURE Q-COLLECT INFRASTRUCTURE

As mentioned above, Collections are numerous and as diverse as the organisms they are hosting. The procedures, the databases and the software they use are very heterogeneous. Some databases are easily useable by third parties while others have extremely limited interest beyond typical cataloguing purposes. Many collections are only useful for their holdings but the associated data are so poor in number and in quality that the viability of the collection is at risk, even if the biological materials themselves are of interest. End-users and clients of collections want more than just a biological store. They want to be able to retrieve associated data, use and analyse them together with other data. Except for a few cases, Collections are unable to provide such advanced services.

In addition, even if the number of biological material available from culture collections is relatively large reaching more or less 2 to 3 million records and hundreds of million for plants and insects worldwide, this number remains an anecdotal portion of the real diversity. Out of the total number of biological material used in scientific papers every year, only 0.01 percent of them are deposited in official culture collections (Stackebrandt E (2010) Diversification and focusing: strategies of microbial culture collections. Trends Microbiol 18:283–287). This means that Collections are far from being used to their full-extent and that many studies are using material that will most likely not be reusable in the future. This is a serious problem. Mobilization of the gathered diversity and associated data is a key factor for the future usefulness and success of Collections.

The future Q-COLLECT infrastructure should contain the following building blocks:

1. A modern, dynamic and efficient collection data management, analysis and publication system
2. Increased acquisition rate of biological material and associated data not stored in Collections
3. Each participating Collection should manage their internal system to be compatible with a common global system
4. Data should be shared and accessible via a central portal allowing researchers and other databases or web services to interact with the Q-COLLECT/EPPO/Q-BANK system
5. Advanced statistical data analysis system available that would take advantage of Q-COLLECT/EPPO/Q-BANK related data and would link them to third parties data
6. Advanced ontological and semantic web technologies implemented in order to allow complex investigations on combined datasets.

To address the issues already discussed above, WP5 members of the Q-COLLECT project propose to create the Q-Book system to be strongly associated, intermixed with Q-bank and EPPO's databases. The Q-Book system would consist in a series of software tools and databases allowing any researcher to gather, store and share data associated with their biological material (not limited to biological material from official Collections). The Q-Book system would allow researchers to perform conventional statistical analyses of data stored in Q-Book/Q-bank alone or in relation to other data coming from other databases or web services. The Q-Book Semantic system would allow to investigate relations between interconnected databases and to retrieve data and properties that could be associated with or linked to biological material.

## 5.1 Q-BOOK SYSTEMS

The basic idea behind the Q-Book systems is to allow any researcher or collection working in quarantine related field to store, manage, analyze and publish data associated to their Biological Material on a freely available platform deeply associated, intermixed with Q-bank and EPPO system. The system would be able to cope with the needs of very small to very large Collections.

The new pipeline will include:

1. A mobile application (IOS, Android and Windows Mobile) to collect sample information (pictures, GIS coordinates, time and miscellaneous metadata) directly in the field.
2. A lightweight desktop application that will synchronize with the mobile app and where a complete and extensible set of data will be recordable such as administration, bibliographical, geographical, ecological, chemical, physiological, medical, molecular, links to other specialized repositories (e.g. GenBank, GBIF, etc.) and many more. This application will not only allow the storage of data but also to perform advanced queries, polyphasic identifications and classifications based on any combination of characteristics.
3. Data will be stored in a local light weight database.
4. Data will be shareable using the central Q-Book facility that will be accessible, searchable by anyone or by a selected list of co-workers.
5. An online or web-based tool will also be created allowing the addition, edition and management of strains, specimens or any biological material data that could be used as an alternative to or complementing the lightweight desktop application mentioned above.
6. The Q-Book website will include user-friendly basic and advanced searching facilities, distribution maps, online polyphasic identifications and web services to download data.
7. A publication tool will also be available to produce e-books or to export data in basic data exchange formats.

Since the Q-Book system would be able to cope with all types of data and organisms, it would be used by a very large panel of researchers and people worldwide. It would be used in many scenarios such as biodiversity management and conservation programs anywhere in the world, quarantine and invasive organisms fight, epidemiology or medical diagnostic. It will be especially beneficial to researchers or collections with no or low financial and technical resources.

There are no such projects proposing to record biodiversity at the unit level (strains, specimens or other biological material). Initiatives such as EOL, Catalogue of Life, Species 2000, IUCN, GenBank, StrainInfo, GCM, BOLD or GBIF are concentrating on species data and are, for some of them, referring to existing and officially recognized museums or culture collections. To our knowledge, there is no complete pipeline that proposes what we are going to initiate here. This ambitious project will be unique and used worldwide. African and lesser-developed countries as well as Collections with no or little resources will certainly be the first to benefit from the proposed system since it will be freely and easily accessible. Most of the African Biodiversity remains unknown and very few African researchers have the ability to buy advanced tools to propose their results to third parties. The Q-Book system will allow them to be visible to the scientific and quarantine problematic community and will inevitably initiate collaborations between research groups that, otherwise, would never know about each other. Since the central database would include a very large range of organisms (from dead material to viruses to higher and large organisms), the diversity and the number of potentially useful data to characterize the biological materials will be huge.

Where possible data deposited in Q-Book will be automatically enriched and linked to existing databases such as Genbank, PubMed or SwissProt (via provided accession numbers) but also, via geographic positioning (latitude,

longitude, altitude, etc) to other datasets like climate, soils, agricultural practices, elevation, vegetation or many other environmental parameters such as demographic, socio-economic or cultural aspects.

Data stored in the Q-Book system will be archived in a coded and a highly structured way and everywhere possible, text data will be avoided to allow descriptive statistics (average, variance, frequencies, etc.) to be easily computed. Each data point will be considered by the system as a programming object and usable by object oriented programming languages.

Such a system will allow the statistical tools to analyse large datasets and perform correlation analyses (for example but many more will be possible as well, like factorial analyses). The semantic system included in Q-Book will allow users to navigate through heterogeneous but highly connected datasets and to find biological material that have potential properties in view of specific Q related research. Statistical and semantic tools will be operable via simple user interfaces from either simple desktop applications or web based systems. This would allow non-advanced users to perform basic calculations, statistics or navigate through complex datasets in a simple way.

The proposed facility will create a virtual research environment where users will be able to find, connect, correct, combine, visualize, analyse and interpret new combinations of heterogeneous data that are relevant to address their Q-research areas. A framework for resource integration will be developed, combining the taxonomic backbone for species identification, geo-referenced environmental data, existing systems for observational data, and information on species traits and attributes. These resources will become accessible to users within seconds, by (i) providing searchable metadata that describe the data and its source, (ii) assigning unique identifiers as linkers between data according to accepted standards, (iii) creating interfaces and query systems, and (iv) installing routines for quality control and interpolation of data in space and time. The architecture is modular, flexible and scalable, and will be the first infrastructure in the world specifically designed for the integration of such Q-resources. Innovative new services will be made available, based on an analysis of the common needs of users in the prioritised areas: services to analyse genetic diversity of mixtures: 'environmental DNA', services for new observational methods using smart sensor technologies, services for constructing species interaction networks, services to link meta-omics data to ecosystem processes in (micro-) communities, for which physical model ecosystem facilities will be created. This being a service-based e-infrastructure means that users can access resources and computational facilities from any place, but also meet in person with technicians when needed to discuss progress and results: facilitating research collaboration is an integral part of the proposed system.

Larger or more advanced collections already having strong CMS will also be able to participate to the project by sharing their data via web services. With their permission, their data will be retrieved through web services and will feed the central Q-Book database allowing to perform queries as well as data analyses on the broadest possible basis (see fig 2).

The Q-Book system will also allow small collections or the ones without an existing data management system to have a complete system allowing to integrate and publish their data using a mobile app, a desktop application or using the web based editor directly accessing their records stored in Q-Book.

The development of the Q-Book systems will certainly not be achieved by the collections themselves since this would require advanced programming and databasing skills that are not or almost never available in current culture collections. Subcontractors with the relevant experience will have to be hired to build the wanted systems in close collaboration with collections experts and Q-bank as well as EPPO, but also with the end-users (existing and foreseeable new or potential ones).

The system will be completely free (free mobile and desktop applications) and the Q-Book system will be hosted and maintained by Q-bank and/or EPPO if the needed resources are made available by one or several financing bodies. The problem of sustainability is one of the major issues that need to be tackled in future projects.

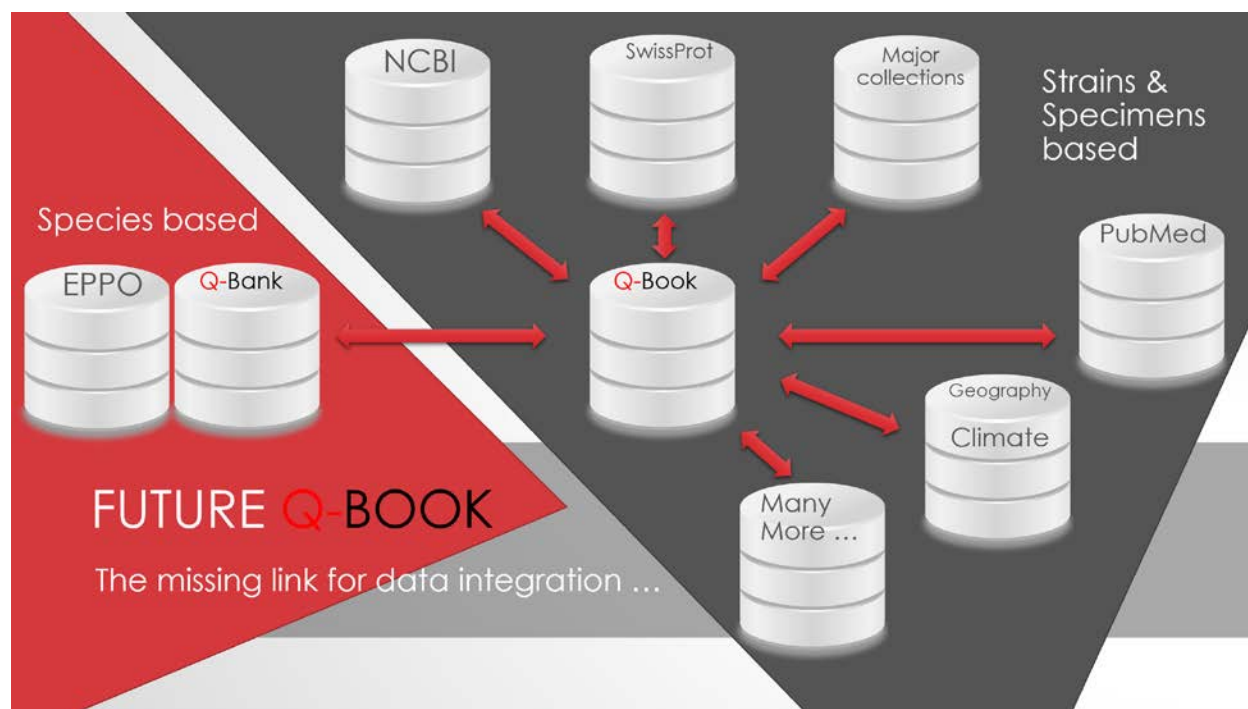


Fig. 1. Q-Book systems & associated systems

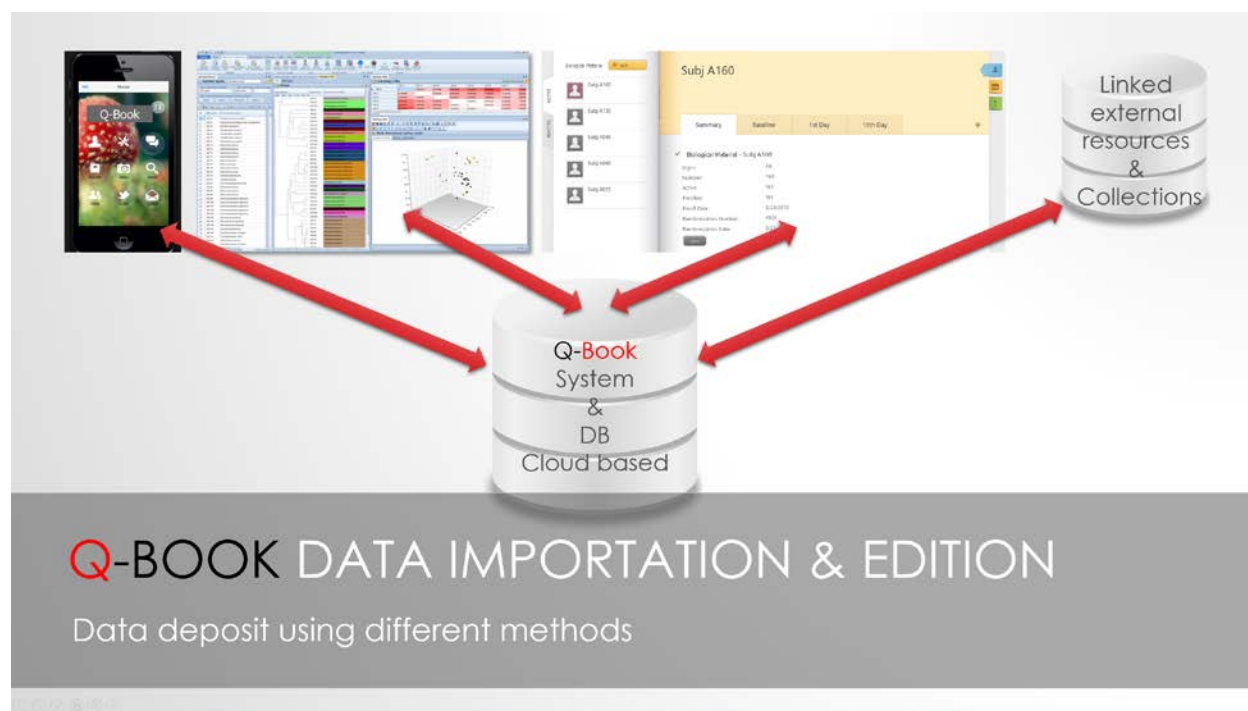




Fig. 2. Q-Book systems for data capture and edition

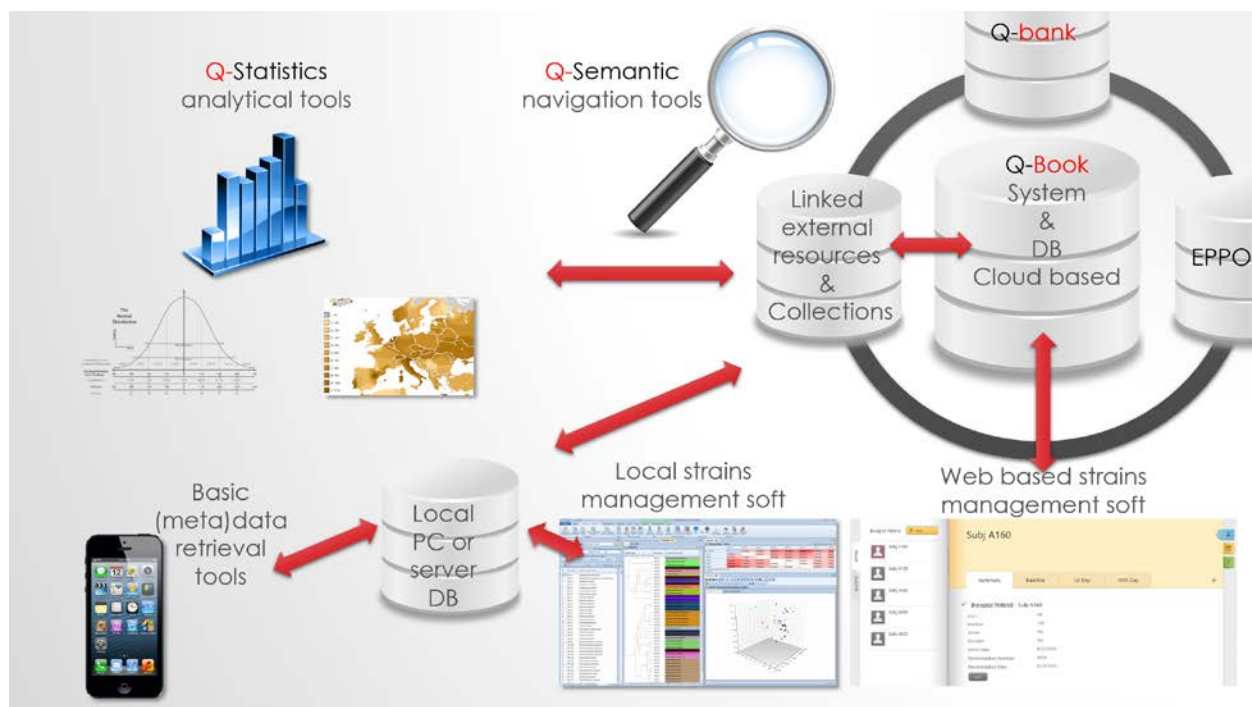


Fig. 3. Q-Book systems functional schema