SYNTHESYS
Synthesis of systematic resources

Report on a generic translator for the communication protocols used in GBIF/SYNTHESYS context.

Work package :	Network Activity D1
Deliverable :	NA-D.1.6.1.
Date :	May 8th, 2007
Author :	Bart Meganck, IT and GIS engineer, RMCA Tervuren
Promotor :	Danny Meirte, curator Herpetology, RMCA Tervuren
Co-promotor :	Patricia Mergen, International Relations, RMCA Tervuren

I. Executive summary

Online availability of high-quality biological datasets is rapidly growing. As scientist become aware of the added value gained from integration of their data in cooperative networks, so the willingness to contribute increases and every month, thousands of new specimen records are added to the providers.

Much work has been done for making the queries on these data providers as easy as possible, and standardising communication and data exchange protocols. Yet multiple incompatible standards exist, each with their own strengths and philosophy : BioCASe, DiGiR, Spice DRM, TAPIR, …

This situation provides flexibility and freedom of choice, but also complicates the software at the user side, as it has to "speak" all the protocols. This report explores the possibility of a generic translator tool for the most frequently used communication protocols, mediating between user and data provider for simplifying interoperability.

The most frequently used communication and data exchange protocols are presented, prior work is discussed, relations with other European and worldwide biodiversity initiatives and further collaborations are explored.

A first, tentative description of a generic translator tool is presented. It aims to be small and simple in its first implementation, with an interface strongly similar to the OGC web service specifications.

II. Citation and license

Meganck, B., 2007. *SYNTHESYS Network Activity D Report Deliverable 1.6.1. Report on a generic translator for the communication protocols used in GBIF/SYNTHESYS context. Royal Museum for Central Africa, Tervuren.*

Table of Contents

III. Introduction

Online availability of high-quality biological datasets is rapidly growing. As scientist become aware of the added value gained from integration of their data in cooperative networks, so the willingness to contribute increases and every month, thousands of new specimen records are added to the providers. The new GBIF data portal currently serves data about nearly 2 million species, from half a million higher taxa, in more than 1000 datasets from 200 countries (http://newportal.gbif.org/welcome.htm , March 6th 2007).

Querying the data in these providers is basically a two-step process : the user sending a query and the provider returning an answer. The answer could be a message related to the communication itself (e.g. an error message), but more often it is just a file transfer with bulk data. All these transactions normally use an XML encoding.

For better focus, this report distinguishes two types of protocols :

the *communication protocols* must be able to establish a connection, and report any failures to do so (e.g. if the service is not available). They don't convey any actual data. Or rather : only the search fields and some variable values (e.g. filter settings).

the *data layout protocols* must carry all the records found in answer to the user's query, possibly thousands of data fields, in a well-defined (possibly hierarchical) structure. In contrast, they needn't occupy themselves with setting up and maintaining connections.

Both protocols still have many varieties and incompatibilities, and most are intimately linked to a particular tool. It is up to the user to make sure that her request is well constructed. As a consequence :

- software systems that want access to the unit data from the providers have to implement all the appropriate communication protocols.
- sending out the same query to two or more servers is generally impossible. This gravely limits the possibilities of cataloguing services, that could otherwise search the web, index accessible servers, and list their datasets.
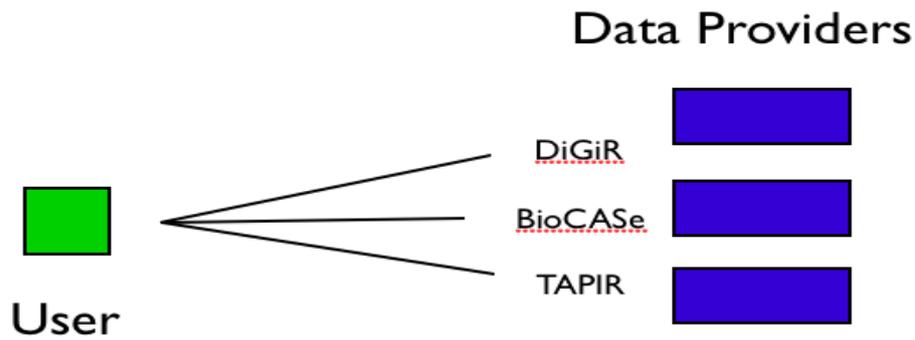
Illustration 1: present situation for search requests : the user software has to implement all possible communication protocols.

III.1.   A generic query translator

To simplify the software at the user's side, a translator system could be considered, hiding the complexity of the different protocols behind a single, simple interface.

The translator would accept user queries in a generic protocol, and pass them on in whichever protocol the data provider expects. The result – possibly converted as well - is returned to the sender.

Ideally, the generic protocol should implement all functionalities present in all protocols it replaces. Technically, it could be an entire new protocol, or an adaptation of an existing one.
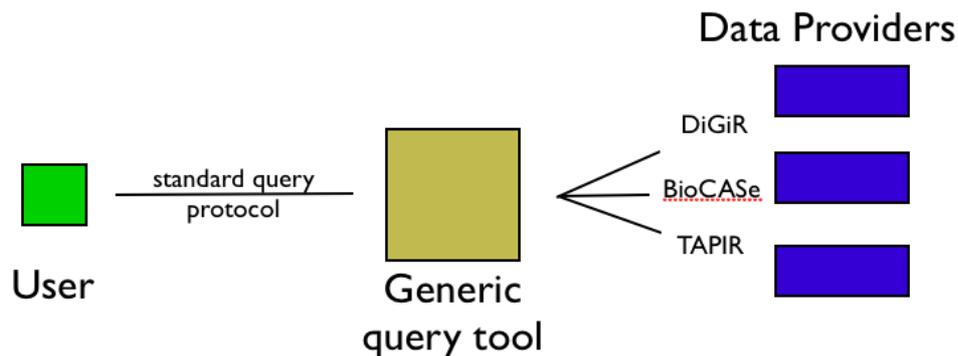
Illustration 2: Search requests through a generic tool. The user software connects with a generic query protocol to the generic query tool. The query is passed on in the communication protocol appropriate for the chosen data provider.

IV. Protocols presently in use in GBIF/SYNTHESYS context

Some important protocols (both communication protocols and data layout protocols), used within the context of on-line distributed datasets (e.g. GBIF, SYNTHESYS, BioCASE initiatives) will now be discussed in more detail.

IV.1. Communication protocols

IV.1.a. BioCASe

(from : *http://www.biocase.org/whats_biocase/index.shtml* and *http://www.biocase.org/products/protocols/index.shtml*)

BioCASE builds on the predecessor projects CDEFD, BioCISE, and ENHSIN. These laid the groundwork for implementing a fully functional service unlocking the immense biological knowledge base formed by biological collections.

The BioCASe protocol was developed within the BioCASE EU-funded project. It is based on the DiGIR protocol, but incorporates some BioCASE specific changes that unfortunately make the two incompatible.

IV.1.b. DiGiR : Distributed Generic Information Retrieval.

(from *: http://www.specifysoftware.org/Informatics/informaticsdigir/*)

The DiGIR protocol was developed by the Biodiversity Research Centre (BRC) Informatics group (University of Kansas - National History Museum) in collaboration with the Museum of Vertebrate Zoology at UC Berkeley and the California Academy of Sciences. DiGIR is currently a public open source project with an international team of contributers, including Centro de Referência em Informação Ambiental (CRIA), Brazil.

DiGiR has been used in many projects and network activities including:

- the Mammal Networked Information System (MANIS)
- the Global Biodiversity Information Facility (GBIF)
- the Ocean Biogeographic Information System (OBIS)
- the ORNithological Information System (ORNIS)
- the global network of herpetological collections data (HerpNET).

IV.1.c. SPICE Common Data Model (CDM)

(from : *http://www.sp2000.org/*)

The goal of the Species 2000 project is to create a validated checklist of all the world's species (plants, animals, fungi and microbes).

The programme is set up in partnership with the Integrated Taxonomic Information system (ITIS) of North America and produces the Catalogue of Life - an Annual Checklist available on the Species 2000 web site and by CD ROM. This is used by the Global Biodiversity Information facility (GBIF) as the taxonomic backbone to its web portal.

For the Species2000 communications, the SPICE Common Data Model (CDM) was developed.

IV.1.d. TAPIR

(from : *TAPIR protocol specification website*)

TAPIR is an acronym for TDWG Access Protocol for Information Retrieval, with TDWG being the Taxonomic Databases Working Group.

The DiGIR (mainly American) and BioCASe (mainly European) protocols have many similarities but are not interchangeable, and this clearly burdens global interoperability.

TAPIR was envisaged as a tool for unifying existing biodiversity data sharing networks based on these two protocols. During development, TAPIR has been refined into a generic product that has the potential to enable interoperability with domains other than biological observations and specimen collections, including geological, ecological, climate, gene sequence, geospatial data and others (by means of extensions). The development of a generic protocol has lead to increasing convergence with similar tools and standards used in a wider context, for instance, the OGC Web Feature Service (WFS) standard (where biodiversity data can be regarded as a geospatial feature).

The new GBIF data portal can index data provided using TAPIR with the DarwinCore schema, and its curatorial and geospatial extensions.

descriptive markup language	descriptive markup schema	Use context
XML	DiGIR	MANIS, ORNIS, GBIF, HerpNET, OBIS
XML	BioCASe	BIOCASE
XML	SPICE Common Data Model (CDM)	Species2000
XML	TAPIR	see DiGIR and BioCASe

Table 1: Selection of communication protocols commonly used in the context of online distributed Biological database initiatives.

IV.2. *Data layout protocols (data standards, data schemas)*

IV.2.a. ABCD

(from : *http://ww3.bgbm.org/abcddocs/AbcdIntroduction*)

The Access to Biological Collections Data (ABCD) data exchange standard was developed by the Taxonomic Databases Working Group (TDWG)/CODATA work group, with one sub-section working on the search and retrieval protocol, **DiGIR**, and a second working on a specification for biological collection data (the ABCD data standard).

There have been many consequent versions of the ABCD protocol. Version 1.2 is currently in use with the GBIF and BIOCASE networks. Further versions improve the compatibility with other protocols, and applicability in multiple fields of science (botany, zoology, bacteriology, virology,...).

The new GBIF data portal is able to search data provided using TAPIR with the ABCD schema, version 1.2, 2.05 and 2.06.

IV.2.b. BioCASE Collection Profile (BCP)

(http://www.bgbm.org/biodivinf/Schema/)

BioCASE Collection Profile is a data schema used in BioCASE to share metadata about collections, organisations and networks. As such, it provides an addition to the ABCD schema, which does not contain information of this level.

IV.2.c. DarwinCore

(from : *http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome*)

The Darwin Core (sometimes abbreviated as DwC) is a standard designed to facilitate the exchange of information about the geographic occurrence of species and the existence of specimens in collections. Extensions to the Darwin Core provide a mechanism to share additional information, which may be discipline-specific, or beyond the commonly agreed upon scope of the Darwin Core itself.

Darwin Core was originally a product of The Species Analyst Project at the University of Kansas. It was proposed as a draft standard for the Taxonomic Databases Working Group (TDWG).

descriptive markup language	descriptive markup schema	type of data	hierarchical profile	Use context
XML	ABCD	specimen data	high	BIOCASE
XML	DarwinCore	specimen data	low	DiGIR
XML	BioCASE Collection Profile (BCP)	collection metadata	high	BIOCASE

Table 2: Selection of data standards commonly used in the context of online distributed databases.

V. Earlier work on protocol and standard interoperability

V.1. Classes definition

Patricia Kelbert and Markus Döring made headway in defining the object classes needed, and constructing a first Python implementation for use in the SYNTHESYS/BioCASE portal

(see http://www.biocase.org/whats_biocase/people.shtml)

V.2. The Biodiversity Users' Framework For Interoperability Experiment (BUFFIE).

(from : Sundar et al.,2005 (1), Sundar et al., 2005 (2))

BUFFIE is a proof-of-concept system for achieving interoperability among various biodiversity data resources. Its main objective is to provide a common platform so that the biodiversity users can access the information that is held in heterogenous data resources, who adapt to different network standards.

BUFFIE is developed by Cardiff University, in cooperation with ETI Bioinformatics, University of Amsterdam, Jerusalem University and Biology Zentrum, Austria.

BUFFIE is an extensive framework with many progressive features. The scope of the generic translator tool presented here is much more modest.

These initiatives will provide a valuable basis for further work, and opportunity for cooperation.

VI. Relation to other GBIF and SYNTHESYS activities

The generic translator tool is of direct importance to current and planned initiatives within GBIF/SYNTHESYS context :

VI.1. The TDWG BioGeo Interoperability Task Group

(from : http://wiki.tdwg.org/twiki/bin/view/Geospatial/GeoAppInter/)

The aim of this task group is the creation of an interoperability framework and possibly an infrastructure where different biodiversity geospatial applications can communicate and produce new results. There are a lot of biodiversity geospatial data providers, from occurrence data to migration routes, distributions, itineraries, etc. There is also several projects working on analytical tools to use this data. But all these systems do not interoperate! Every system is created as a kind of silo, and the consumption of its data or results is not done in a very efficient way. By using OGC standards the Task Group aims at solving the interoperability problems among these services and move forward to a Biodiversity Geospatial Web where all services can be chained in a workflow.

The generic translator tool will be extensively discussed within this Task Group, as it is perfectly aligned with its mission of improving interoperability.

VI.2. SYNTHESYS Network Activity D 3.7 "itineraries" project

The "Itineraries" project reconstructs an expedition's travelling route from the georeferenced gathering data provided by the user. She can upload her own data in a text file, and/or send a query command to an online data provider like GBIF. The raw data (georeferenced specimen records) are searched by the itinerary algorithms, and the most probable journey pathways are displayed on the web (via Web Map Service). With this visual aid, inconsistencies within the dataset can easily be spotted. The tool thus serves for visualisation and data quality control.

The generic translator tool will greatly simplify the input mechanisms needed for the "itinerary" implementation. The tool is therefore developed as an additional task within this project.

VI.3. EDIT WP5 Internet Platform for Cybertaxonomy

(from : http://wp5.e-taxonomy.eu/)

Work-Package 5 (WP5) An Internet Platform for Cybertaxonomy is one of the focal points of EDIT as it is directed at the practical application of methods developed in information science and biodiversity informatics for use in revisionary taxonomy and taxonomic field work. At the same time EDIT hopes to facilitate the integration of biodiversity informatics and IT resources within participating institutions. In conjunction with these efforts, it will reach out to identify existing electronic tools for taxonomists, test their usability and where

necessary provide developer time to improve their interoperability.

The aim of EDIT is to build a distributed computing platform that assists taxonomists by providing certified and tested existing labour and time saving tools in order to do taxonomy expediently and via the web.

The integration envisaged within the EDIT initiative will greatly benefit from each tool that makes for better interoperability.

VII. Collaboration on the development of the generic translator tool

Development of this tool was planned for spring 2007. In GBIF/TDWG context, this perfectly coincided with the TDWG GeoInteroperability Testbed Pilot Project (http://www.tdwg.org/homepage-news-item/article/tdwg-geointeroperability-testbed-pilot-project/) : a hands-on workshop about protocols and interoperability of tools in the context of networked biological databases. The insights and prototypes from this meeting will assure that the generic translator tool is well aligned with the other interoperability initiatives.

As close collaboration is paramount for interoperability, a two weeks' working visit to BGBM Berlin is also planned. The know-how and experience of the people that helped develop the standards and protocols of GBIF are invaluable for this work. We envisage an technical and conceptual cooperation period of about two weeks, with the following schedule :

•   Getting acquainted, preparatory exploration of ideas in face-to-face dialogue. Expectancies and mindsets of both sides. Alignment of concepts and definitions, to avoid misconceptions that could burden the cooperation

•   introductory presentation to the Berlin colleagues of our work at RMCA, the national and international synergies of our work, the projects we are working on.

•   overview of the BGBM collection, and its concept representation in the database. Discussion of differences and similarities in comparison with the RMCA collections and database layout.

•   quick exploration of prior work done by both parties : partial proofs of concept, tentative implementations, existing code. Exchange of ideas on working tools, best practises,tips and tricks for programming.

•   discussion on priorities for technical implementation of a prototype : discussion on the functions to include and those not to include. Discussion on test cases. Documentation of agreed conclusions.

•   setting up schema for unit testing, loading of test datasets.

•   development of prototype. Actual code writing, starting from a nucleus

with basic functionality and gradually adding more features.

- documentation of discussions and decisions. Technical and functional documentation.

Other collaboration opportunities are actively explored.

VIII. Definition of service

VIII.1. General layout of the service

The generic translator tool is meant to simplify the connection process. Its use should be as transparent as possible. At the user side as well as the provider side, it should require the least installation or configuration possible.

OGC webservices (WMS, WFS, WCS and the like) set a good example in this respect. They offer a functionality that is broadly similar to what the generic translator will provide : a web-based service with a minimalistic protocol and standardised (XML) response.

The recent TDWG BioGeo Interoperability Task Group Meeting (Campinas, april 2007) demonstrated clearly the ease of use of the OGC webservice tools in an integrated workflow :

- the variables (key-value pairs) are visible in the URL string, which makes for easy debugging and adaptation for quick tests.

- the URL GET method is well-known to administrators and programmers.

- people in the field of biodiversity informatics are familiar with the OGC methods and interface.

- results from the service (data results or error messages) come back as standard XML.

We therefore chose to implement the generic translator tool in a similar way, using the same request parameters where appropriate. The generic translator tool does NOT, however, claim to be a valid OGC webservice. At most, we use more or less the same syntax for easy recognition and, perhaps later, integration.

VIII.2. Provisions for low-bandwidth environments

Not all institutions involved in biodiversity research have high-speed access to the Internet. The Royal Museum for Central Africa, with its many contacts in the African continent, is well aware of these issues.

At times the connection is very limited, or expensive, or only available during a certain time of day. In such cases it would be interesting if the protocol could adapt to sparser settings, for example by providing quick previews (e.g.

only a few answers), or by compressing result datasets before transfer.

VIII.3.   Overview of the protocol

A first, tentative, description of the basic protocol as we envisage it. The aim is to develop a rapid prototype, as small and simple as possible. A first implementation should do some useful work with no more than 2 methods.

VIII.3.a.   GetCapabilities

First and foremost, the translator tool should be able to return some metadata about the data server it will connect to.

This is a variant on the OGC "GetCapabilities" request, which is often seen as one of the most useful features. One important difference, however, is that our service (the translator tool) does NOT return metadata about itself (like OGC webservices do) but only about the server it links to. From the user's point of view, the translator service is completely transparent.

Perhaps in some later stage, the possibility to query the translator service directly could be added. That would make it even more "OGC-like", though more like WFS-T (WFS-transactional, where besides the web service there is the database server to connect to). But this would add a complex duality in the requests : are we querying the tool, or the data server it links to ? Technically, this could well be solved, but for a first prototype it is certainly not yet relevant.

For sending this query, we just need to provide a single parameter : the URL of the data server we want to contact :

http://someserver.somewhere.org/generic_tool.php?

REQUEST=GetCapabilities

&SERVER=http://dataserver.someplace.org/

The response should include the server's metadata, such as is provided in their protocols : which protocol is used ? which version ? who is the contact person for the dataset ? what is the return data schema ?

VIII.3.b.   GetSpecimens:

Crucially, the distant data servers have to be queried.

The GetSpecimens request caters for this. It takes a search string with the botanical name of the species that is wanted. Wildcards should be permitted for searches.

http://someserver.somewhere.org/generic_tool.php?

REQUEST=GetSpecimens

&SERVER=http://dataserver.someplace.org/

&SEARCHSTRING=Phragmites austr%

&MAXCOUNT=500

&OVERVIEW=true

The MAXCOUNT and OVERVIEW parameters implement the possibility to return only a subset and/or preview of the total data set – for quick validation or searches.

The response will be the XML data file provided by the data server in the format he uses (ABCD, DarwinCore,...). In the first instance, this file will be passed on to the user "as such". Later on, some kind of "translation" process could be added – though a full automated conversion of any server given format to any other desired format, seems impracticable.

With no more than these two basic methods, the tool would already be useful. We should not forget that these "simple" requests by themselves represent an extensive body of logic, as all their functions have to be translated into 3 or more other protocols, including the required connection parameters and previsions for error handling. Moreover, the most frequently used transactions are covered : getting the server metadata, and querying the specimens.

VIII.4. *Choice of technology*

In line with the general philosophy governing all GBIF/SYNTHESYS applications, it is clear that the prototype of a generic query tool should be implemented in an open and platform-independent technology. After consideration, PHP server-side scripting was chosen, considering its advantages :

- platform independent : works on Linux, MacOS and other systems
- libraries : many libraries work out-of-the-box : XML, DOM, Unicode
- open-source
- object-oriented

PHP was also used with good results in the TDWG BioGeo Interoperability Task Group Meeting (Campinas, April 2007), where an integrated workflow of OGC-compliant services was constructed.
See : http://wiki.tdwg.org/twiki/bin/view/Geospatial/InteroperabilityWorkshop1

IX. Timeline for further work

• **June 2007** : working visit to BGBM, Berlin. Here, the first prototype of the tool will be implemented, in cooperation.

• **July 31st ,2007** : deliverable NA-D.1.6.2.: working prototype implementation of the service.

• **Fall 2007** : integration of generic tool in RMCA's NA-D 3.7 "itinerary" framework. The SYNTHESYS Network Activity 3.7 will yield an online tool for retracing expedition itineraries based on the georeferenced specimens from online databases such as GBIF. More information on the scope and functionality of this project in Meirte et al.,2006, and Mergen, 2006. A pdf version of the report can be downloaded on the [SYNTHESYS NA-D 3.7 "itineraries" report](#) site.

•

X. references to the Internet :

All hyperlinks used in this report are available hereunder in text format (alphabetically), for use in printed copies of this work.

ABCD :

http://www.bgbm.org/TDWG/CODATA/Schema/

BGBM XML schema repository :

http://www.bgbm.org/biodivinf/Schema/

BioCASE website – what is BioCASE ? :

http://www.biocase.org/whats_biocase/index.shtml

BioCASE website : the BioCASE access protocol :

http://www.biocase.org/products/protocols/index.shtml

BUFFIE : Design and Architecture for the prototype ENBI common data access system – BUFFIE (Biodiversity User's Framework for Interoperability Experiment). ENBI Report No: WP9_D9.2b_04/2005 :

http://circa.gbif.net/Public/irc/enbi/comm/library?l=/enbi_reports/wp9_d9_2b_tr2_pdf/_EN_1.0_&a=i

DarwinCore :

http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome

GBIF data portal :

http://newportal.gbif.org/welcome.htm

HERPNET : http://www.herpnet.org/

MANIS : http://manisnet.org/

OBIS : http://www.iobis.org/

ORNIS : http://olla.berkeley.edu/ornisnet/

SYNTHESYS Network Activity D :
http://www.synthesys.info/network_activities_d.htm

SYNTHESYS NA-D 3.7 "itineraries" report :
http://www.synthesys.info/NA_Documents/Deliverables/NA%20D/SYNTHESYS_milestone_report_may31.pdf

TAPIR protocol specification website :

http://www.tdwg.org/dav/subgroups/tapir/1.0/docs/TAPIRSpecification_2006-10-15.html#toc4

XI. References to literature :

Berendsohn, W. G., 2001: CODATA Working Group on Biological Collection Data Access. A joint CODATA and TDWG initiative. - CODATA Newsletter 82: 7-8.

Berendsohn, W. G., 2002: BioCASE - A Biological Collection Access Service for Europe. Alliance News 29(6): 6-7.

Berendsohn, W.G., 2005, ABCD – the proposed standard XML schema for Access to Biological Collection Data. Abstract on http://www.tdwg.org/2005meet/TDWG2005_Abstract_5.htm

Krumenacker, A. and Malicky, M., 2004. Report on proposed data standards and protocols with respect to analysis tools. Work Package 10 : generic analysis tools and data mining. ENBI report No. WP10_D10.1b_08/2004

Meirte, D., Mergen,P., Meganck,B. and Theeten, F.,2006. SYNTHESYS NA-D 3.7 Providing itinerary related datasets and tools (for integration, visualisation and quality check).

Mergen, P., 2006. Mission Report, SYNTHESYS NA-D 3.7 Providing itinerary related datasets and tools (for integration, visualization and quality check)

Sundar, R., Gray, A., Jones, A., White, R., Brugman, M., Addink, W., Kaletas, E., Beller, A. and Krumenacker, A., 2005 (1) . Design and Architecture for the prototype ENBI common data access system – BUFFIE (Biodiversity User's Framework for Interoperability Experiment). ENBI Report No: WP9_D9.2b_04/2005. Cardiff University, School of Computer Science, UK. ETI – BioInformatics, University of Amsterdam, the Netherlands. UVA, Amsterdam, the Netherlands. Jerusalem University, Israel. Biology Zentrum, Austria.

Sundar, R., Gray, A., Jones, A., White, R., Brugman, M., Addink, W., 2005 (2). Feasability study for prototype common access system. ENBI Report No: WP9_D9.3a_12/2005. Cardiff University, School of Computer Science, UK. ETI – BioInformatics, University of Amsterdam, the Netherlands.