Milestone report for SYNTHESYS Network Activity NA-D : Developing and maintaining databases.

NA-D 3.7

Providing itinerary related datasets and tools for integration, visualisation and quality check - system specifications -

Contractor : Royal Museum for Central Africa (RMCA), Tervuren.

Date : may 31, 2006

Authors :

Bart Meganck, IT engineer

Danny Meirte, promotor

Patricia Mergen, co-promotor

Franck Theeten, IT engineer

Executive summary

The SYNTHESYS 3.7 "itinerary" project is developed within the broader framework of biodiversity informatics. Itineraries use the knowledge about expedition pathways to detect errors or anomalies within the expedition dataset. Existing tools for data quality assessment focus primarily on separate data points, whereas itinerary tools consider the coherence of clusters of data.Thus, they will be complementary to prior and ongoing activities in the field.

Knowledge about the expeditions can be obtained from various data sources, both digital and not (yet) digital : field diaries, hand-drawn maps, specimen database records, written comments, rough terrain sketches, digital maps, field number lists and others. The possibilities of these sources were explored in close cooperation with our SYNTHESYS partners, and by synergies with other projects in the field, leading to the conclusion that such sources are both available and usable.

As a first analytical step, a formal Unified Modelling Language (UML) description of the concept of "itineraries" was presented, providing a framework to fit all entities within the itinerary project. This schema describes the terms used ("expedition", "itinerary", "event"), their definition within the project, and their mutual relations. This greatly helps in using a common terminology, developing a standard for itineraries. This prior analytical work will serve for the implementation and visualisation phase.

The itineraries description dataset has been developed as an additional concept to the official Access to Biological Collections Data (ABCD) standard of the Taxonomic Databases Working Group (TDWG). Thus already existing ABCD concepts useful for itinerary description have been identified and if needed additional ABCD concepts for itineraries can be submitted for the next version of the ABCD standard. The identification of the fields made clear which information should be provided to the itinerary tools, and how it can be standardised.

Central to the whole of itinerary tools is the algorithm for deciding which data are consistent with a given itinerary, and which data are not. Various approaches for constructing such an algorithm were explored and their pros and contras assessed. Subsequently, a prototype for an algorithm was proposed and discussed. When implemented, it will filter out relevant data records and flag possibly erroneous ones such as outliers - the very aim of the itinerary concept.

Interpretation of the complex data sets from the expeditions is facilitated by various methods of visualising. This will be the primary form of interface between the end user and the itinerary tools. Different possibilities were tested, and examples are presented.

Acknowledgements

This report has received the generous help and support from many people and institutions.

In particular, we would like to thank :

* Danny Meirte, promotor, for initialising and inspiring the "itinerary" idea, for daily support, knowledge and experience sharing, and for the regular idea exchanges needed for keeping focussed.

* Patricia Mergen, co-promotor, for her valuable support in communications about the project, for background documentation and tips, and daily support.

* Franck Theeten, IT-specialist, for technical advice, tips & tricks.

* Our SYNTHESYS partners, for their cooperation and response : Walter Berendsohn, Markus Döring, Anton Güntsch, László Peregovits, Javier de la Torre.

* lat/lon GmbH, Bonn, for the DeeGree software support and development.

Furthermore, many people were prepared to kindly provide their own datasets for internal testing :

* Ugo Dall'Asta, RMCA, Lepidoptera.

* Marc de Meyer, RMCA, Diptera.

* Rudy Jocqué, RMCA, Arachnea.

* Michel Louette, RMCA, Aves.

* Danny Meirte, RMCA, Amphibia, Reptilia.

* Jos Snoeks, RMCA, Ichtyology.

Several SYNTHESYS partners have already responded to our request for collector information. This information will serve for testing and implementing the algorithms and the visualisation, starting June 2006 :

* The National Herbarium of the Netherlands, The Netherlands.

* The Zoological Museum, University of Copenhagen, Denmark.

* The Naturhistoriska Riksmuseet, Stockholm, Sweden.

* The Hungarian National History Museum, Budapest, Hungary.

The SYNTHESYS project was made possible by funding of the European Union.

Our most sincere apologies should go to any persons or institutions who helped but are, despite our best care, not mentioned here. Should this occur, we would be very obliged to hear from them and to make corrections.

☑*note : hyperlinks :*

☑ All URLs and hyperlinks embedded in the electronical version of the text are given in full writing in the references, alfabetically sorted.

Table of Contents

I. SYNTHESYS itinerary project : introduction and background.

I.1. SYNTHESYS and the itinerary project within the biodiversity informatics context.

I.1.a. Biodiversity informatics

The Global Biodiversity Information Facility (GBIF) and the Biological Collection Access Service for Europe (BioCASE -see Berendsohn, 2002) initiatives serve millions of biological records through their transnational networks, unlocking vast resources of knowledge for scientists in standard formats and protocols promoted and endorsed by the Taxonomic Databases Working Group (TDWG) and the Committee on Data for Science and Technology (CODATA), such as :

Standards for data encoding :

* DarwinCore

* Access to Biological Collection Data (ABCD)

* Structure of Descriptive Data (SDD)

* Taxonomic Concept schema (TCS)

Standards for data exchange :

* Distributed Generic Information Retrieval (DIGIR)

* TDWG Access Protocol for Information Retrieval (TAPIR)

* Biological Collection Access Services (BioCASE)

Interest in georeferencing of these records is quickly mounting. Increasing concerns about endangerment of economically important services provided to humanity by biodiversity have led to increasingly urgent calls for good scientific data and information on which to base management decisions (http://www.gbif.org). New and user-friendly graphical tools become valuable assets in interpreting complex data patterns and monitoring changes in fields ranging from biodiversity assessment to collection management.

I.1.b. The SYNTHESYS project

The European Synthesis of Systematic Resources (SYNTHESYS) project aims to raise awareness of best practices in this matter by offering improved training and workshop opportunities, and guidelines for the care, storage and conservation of collections (http://www.synthesys.info).

In all this, the quality of the primary data is of the utmost importance, as it is the raw material from which all further assumptions and conclusions are derived. It is an explicit objective of the SYNTHESYS Networking Activity D to increase the technical quality and availability of networked data resources (http://www.synthesys.info/network_activities.htm).

Itineraries provide a means to achieve that purpose.

I.1.c. Itineraries

Many of the georeferenced data available have been collected during expeditions and

surveys. Therefore, knowledge about these journeys can be used to detect errors or anomalies within the dataset. From Danny Meirte, curator Herpetology in the Royal Museum for Central Africa (RMCA), came the idea and the proposal for using "itineraries".

I.1.d. The concept and scope of itineraries

(after Meirte, 2005) :

* To detect itinerary patterns in georeferenced primary data presumably collected during a collecting event. A first validation approach is to use georeferenced primary information from well-known itineraries and to evaluate if itineraries obtained from the coordinates and collecting date correspond to what is known from literature.

* In a second step, the defined algorithms will be tested and applied to georeferenced primary data available in the GBIF and BioCASE network in ABCD and DarwinCore, where the expeditions route are less documented or even completely unknown. It is likely, depending on the accuracy of the available data, that several possible alternative expedition routes will be extrapolated. These routes and the related collecting points will be shown to the end users on on-line maps using GIS services. These latter tasks will be done in close collaboration with SYNTHESYS NA 3.6. The same tools and data standards will be used.

I.1.e. Itineraries : additional data sources.

Some itineraries which were followed by the collecting campaigns are well known for historical reasons (i.e. famous expeditions in the Belgian Congo or in Polar regions), but many others are only poorly known or documented (Mergen, 2006). Therefore, additional data sources can be considered, even if these are not (yet) available in electronic form :

* field notebooks
* hand-drawn maps
* written commentaries on maps
* rough terrain sketches
* field number lists
* specimen labels
* ...

Inquiries with the RMCA collection managers, and a networking visit to the National Botanic Garden of Belgium (NBG), made clear that many interesting sources of such information are indeed available.

I.1.f. Additional information harvest.

When the itinerary tools detect errors or inconsistencies, some of these may be automatically corrected. Others can be listed - with suggestions of possible causes - for closer (human) examination. Parameters steering this process can be adapted. With experience, the best parameter values for a given dataset will become apparent. Relevant external information (e.g. from digitised field journals) may be added - if available - for additional precision.

Ultimately, the parameter settings will provide valuable information about itineraries, beyond the basic task of data quality control :

* the most likely pathways between localities in a certain time
* the average speed of various means of transport
* field number characteristics for certain persons or institutions

* common references to certain regions

* recurrent inaccuracies on certain maps

* alternative calendars in use...

Such additional information could in turn increase the performance of the itinerary tools. The extraction and storage methods for such information, however, lie outside the NA-D 3.7 project scope. It should be mentioned here, though, that a very similar kind of information gathering is currently being explored by Markus Döring and Anton Güntsch in the annotation services system as described in Mergen,2005(2) in the framework of BioCASE, GBIF-Germany and other SYNTHESYS NAD tasks.

I.1.g. Synergies with other data quality tools.

Many data cleaning and checking tools are already in use. The BioGEOMancer and DeeGree projects have various initiatives for data validation, outlier detection, and data cleaning software and best practices. Data providers for the GBIF nodes can use the GBIF Data Tester , developed by GBIF in collaboration with the Centro de Referência em Informação Ambiental (Brasil). Amongst others, these tools check for :

* unrecognised values for data elements (e.g. country names or basis of record values).

* coordinates falling outside the boundaries of named geographic areas.

* scientific names that are not known to external lists such as the Catalogue of Life or nomenclators.

* appropriate formatting.of scientific names

The itinerary project does not intend to duplicate these efforts, but acts as a complementary tool set. In particular, it will focus on such anomalies as:

* wrong georeferencing caused by name ambiguity

* nominal collectors on the specimen's label

* "neutral" typos on specimen labels, meaning typos that make no difference to a simple existence check, e.g. if both 'Altman' and 'Altmann' would exist as explorers.

* omitted data on specimen labels

* specimens wrongly assigned to an expedition

Both lists of possibilities are not exhaustive, but the difference in emphasis will be clear. The itinerary tools will work on a level closer to metadata, on errors more related to whole clusters of data, than to separate data points. Thus, there will be synergies with existing tools.

I.1.h. Collaboration within the SYNTHESYS itinerary project.

The SYNTHESYS itinerary project work is done in close collaboration with the SYNTHESYS Taxonomical Facilities (TAFs), and builds on prior work of standardisation bodies in the field of Geographic Information Systems (GIS) like the Open Geospatial Consortium (OGC) and TDWG in the field of taxonomical and biodiversity related information.

The technical basis has been prepared by the European Network for Biodiversity Information (ENBI). Work Package WP-10, produced two reports on software tools and standards :

WP10,D10.1a, Report on available analysis tools and proposed choices for ENBI network (Beller *et al.*, 2004).

* gathered and reviewed 759 different analysis software packages in 5 different categories.

* named and described amongst others Diva-GIS, DeeGree and PostGIS, which are now used in the project.

WP10, D10.1b, Report on proposed data standards and protocols with respect to analysis tools (Krumenacker & Malicky, 2004).

* Went more in depth on some of the software of the WP10-D10.1a report, most notably DeeGree WMS, and its use of GML.

SYNTHESYS NA-D 3.6 Deliverable 3.6.1 : Report of existing GIS standards and software (Torre, 2005).

* The relation between geospatial information projects related to GBIF was mapped in an architecture proposal. Here, links and integration of itinerary services, spatial servers, data providers and existing projects was explored (illustration 1).

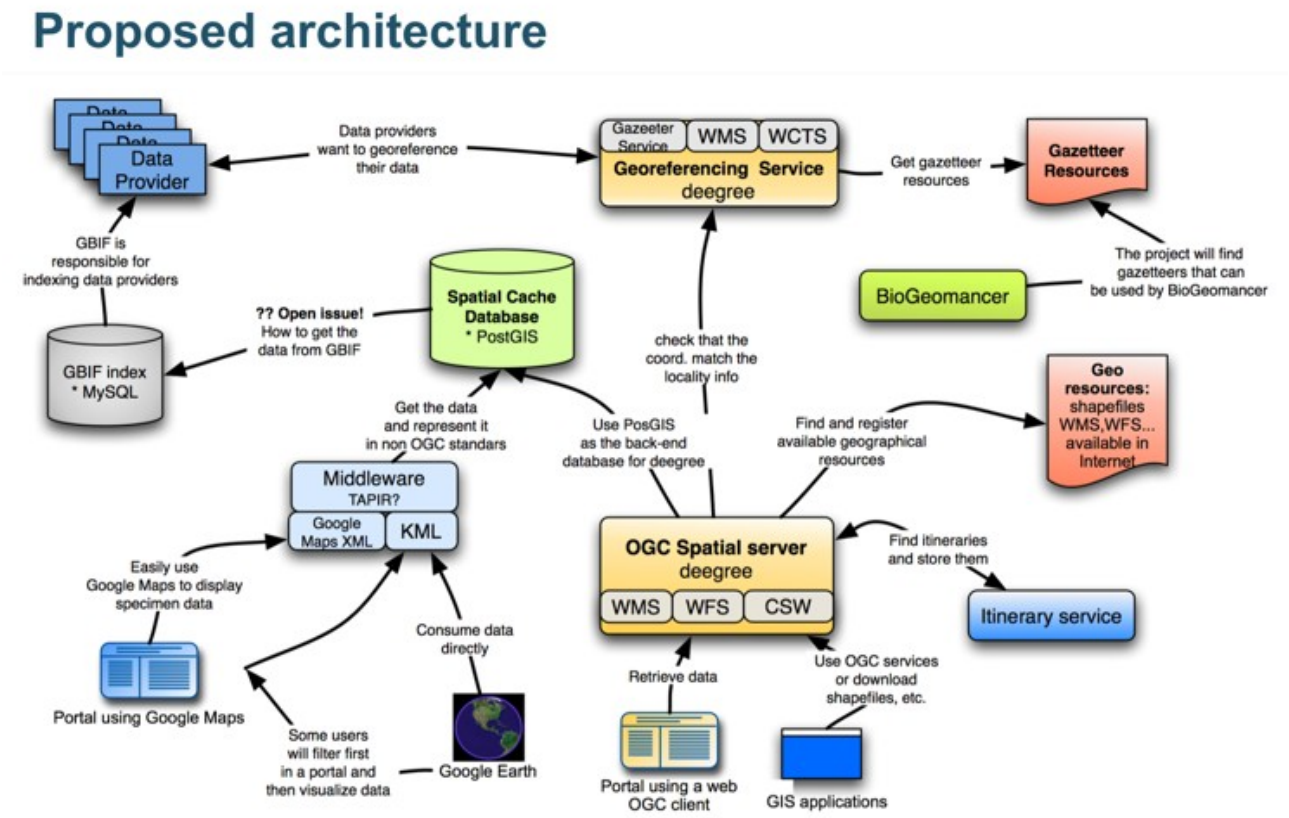


Illustration 1: relation between geospatial information projects related to GBIF. from: Torre,2005.

Several SYNTHESYS-specific meetings coordinated the efforts between the partners, and provided requirements for the Graphical User Interface (GUI), GIS standards and software :

SYNTHESYS NA-D 3.6-3.7 and GBIF.DE Co-ordination meeting, Bonn, October 2005 (Mergen, 2005 (1)).

* The possibility of using Google Maps and Google Earth / KML to display data was explored. The itinerary project will display its data in Google Earth (amongst others).

* The schema of relations between geospatial information projects related to GBIF was proposed (see : Torre, 2005). This schema has been a guide for collaboration within the itinerary project.

* The possibility of including lat-lon GmbH in the developments was considered. Lat-lon is now actively participating in meetings and in the thinking process.

SYNTHESYS NA-D User Interface task force meeting, Paris January 2005

* made an overview of subtasks and responsible institutions for implementation. This produced the appointment of task which we now follow.

* made a timetable for subtasks and deliverables. This timetable is now in use to time and coordinate our efforts.

SYNTHESYS NA-D User Interface meeting, Budapest, November 2005 (Mergen,2005(2)).

* constructed a draft plan and schedule for the RMCA itinerary work.

* coordinated tasks so as to prevent duplication of efforts.

Synergies between current projects are actively explored, and key players have been bought together for exchanging ideas. In particular :

synergies with the implementation of the visualisation server for spatial unit data, and the web feature service and query system within NA-D 3.6, by Consejo Superior de Investigaciones Científicas (CSIC) Madrid.

consulting of the lat/lon company has been obtained for the deployment of the DeeGree geospatial server. This software is "the most substantial implementation of OGC- and ISO-standards in the field of Free Software". Ideas for optimisation of DeeGree with respect to the itineraries project are relayed to lat/lon, and actively discussed for feasibility, so this "consulting" is much more than just a one-way relationship. For the contact with lat/lon and our partner institutes, a special GIS-GBIF mailing list (gis_gbif@yahoogroups.com) has been created. It covers all GIS-related questions and proposals.

* The SYNTHESYS TAFs are involved in the quest for relevant (testing) data, and gathering information about external information sources (be they digitised or not) useful for further development within the itinerary project :

A visit was paid to the National Botanical Garden (NBG) of Belgium TAF in Meise. The visit showed the potential of "hidden" (i.e. not yet digitised) material for the further development of the itinerary project. It was very clear that lots of interesting data could be found in field diaries, written lists, old maps, handwritten notes and sketches.

A request for information overview has been sent, by email, to all TAFs involved in SYNTHESYS, asking for any data that could be relevant for the itinerary project. Answers were received and if necessary additional information was provided. Collaborations and test data were considered in mutual correspondence. A comparison of the information provided will be made to detect possible synergies in ongoing projects and the possibility of joint publications.

* The RMCA was host for a workshop on GIS related projects around GBIF and TDWG, encompassing the itinerary work and related projects (see below).

I.1.h.1. The RMCA workshop on GIS related projects around GBIF and TDWG.

On February 22, 2006, the RMCA hosted a workshop in the framework of the SYNTHESYS itinerary project.

Participants were colleagues from our SYNTHESYS partner institutes, German and Spanish GBIF nodes, staff of the lat/lon Company, GNOSIS project and RMCA staff from different disciplines :

* Javier de la Torre (CSIC, Madrid, task leader of Core GIS services - SYNTHESYS NA-D 3.6)
* Jesús Fernández Segovia (Real Jardín Botánico de Madrid, GBIF Spain)
* Ramón Pérez (Real Jardín Botánico de Madrid, GBIF Spain)
* Steven Smolders (GIM, Belgium, GNOSIS project)
* Christian Kiehle (Geographisches Institut, University of Bonn)
* Jorg Holetschek (Botanical Garden, Berlin, University of Berlin, GBIF Germany)
* Pascale Lahogue (RMCA Geology Department)
* Michel Louette (RMCA, African Zoology Department)
* Danny Meirte (RMCA, African Zoology Department)
* Patricia Mergen (RMCA, African Zoology Department)
* Bart Meganck (RMCA, African Zoology Department)
* Franck Theeten (RMCA, African Zoology Department)
* Garin Cael (RMCA, African Zoology Department)

First, different talks gave an overview of the current status of the various projects :

* Current Status of the NA-D 3.6 work in Madrid. (Javier de la Torre)
* Itinerary services + demo of existing visualisations (Bart Meganck, Patricia Mergen). Here the RMCA presented the context of the itinerary services, the concept of itineraries, the work that had been done, and the further steps that would be taken.
* GNOSIS Demo (Steven Smolders)
* GBIF Spain, RMCA Geology, Botanical Garden : status report and open discussion.
* GBIF Germany and lat/lon : status, possibilities of DeeGree software.

Next, there was an open discussion :

* Exploring synergies between existing GIS related projects
* Compiling the user requirements

Finally, the future work was presented :

* Tasks remaining in the ongoing projects
* Plans for future projects.

In relation to the itinerary project, amongst others the following items have been discussed:

* The availability of gazetteers, and the possibility to create an African gazetteer
* Use of itineraries for georeferencing
* UML diagram of itinerary concept
* Different possible representations of features and feature types
* Overlaying itinerary WMS layers on the Madrid interface
* KML output format for DeeGree
* Converting GML to KML

The announcement and complete agenda of the workshop can be found in annex.

II. Problem analysis

Detecting itinerary patterns within the primary data (specimen records) is a task that requires a well-structured approach. There are currently some 90 million data records in the GBIF domain alone (http://www.gbif.org/DataProviders/statistics1), and these figures are increasing continuously.

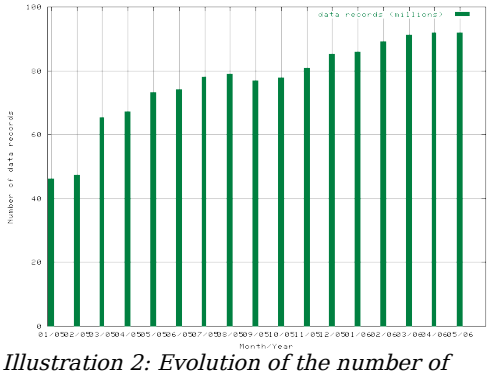


Illustration 2: Evolution of the number of data records in GBIF providers.

That is a vast dataset, offering different kinds of information. For the itinerary construction, the most relevant information is :

* temporal information
* spatial information
* information about participants

Even if all this is available, an itinerary description can not readily be distilled. As in statistics, only "model" well-known itineraries are used for setting up a modelling algorithm and to test the model for validation. Once the model is tried and tested, an assessment can be made of which data fit in. It makes good sense to start with a modest, well known dataset that can be easily controlled.

II.1. Approach

II.1.a. Defining a framework.

Even in well-known datasets, some data points are less suited than others for defining a model, due to the greater uncertainty about them. For example specimen collection points are usually less well-defined :

* the party is often still under way (spatial uncertainty)

* the party can be split into different sub groups for the day (participant uncertainty),

* extensive notes are often only made in retrospect, i.e. at the night camp or village (temporal uncertainty).

Other points could be more useful :

* they are better defined in time and space.

* they can be correlated with other information, for checking.

For example, spatial information offers a wide range of possible correlations :

* methods of georeferencing

* relation to other localities

* former names of localities

* relation to means of transport

In particular, places were the night was spent are provide excellent hooks to attach all other activities and events. These places tend to be well-defined :

* in time, because you have (normally) the midnight hour with some hours before and after.

* in space, because in most cases it will be a well-marked locality (e.g. a village).

* in correlation : the sequence of camping spots is one of the things a field diary or an official report mentions very frequently.

* in participant composition

Thus, it would make sense to construct the basic framework for our itinerary description from these camping points. All other data points, including specimen collections, could then be attached to this framework.

Every one of those additional points will either confirm this basic outline, or pose a contradiction, and this is what the itinerary algorithms will detect.

Everything together will yield the eventual itinerary description.

II.1.b. From framework to itinerary description.

Next will be proposed the work flows for constructing an itinerary description from the available information :

* starting from literature for well-known expeditions (thus working top-down), setting up the model and algorithm prototypes.

* starting from the unit-level data for little known journeys (thus working bottom-up), after the model has been validated and tested on unit-level data.

II.1.b.1. Working top-down

For a well-known expedition, literature review provides a lot of information.

From this information, two documents are made :

* a description rule set, being a formal description of the framework of the expedition (as described above).

* a conformity rule set, being a set of values for parameters to decide whether additional points are conform with the itinerary description rule set (for example, a maximum speed of travel for a given means of transport).

A conformance check takes in the unit-level dataset, and decides which records correspond to the conformity rule set. These records are passed on, others are ignored.

The records that are passed on, are checked for their possibility to constitute new events. :

* either the information they provide improves the general description framework of the itinerary, and they should be added to the description rule set

* or the record, though consistent with the description, does not add new information. In that case, it is added to the dataset of the additional points.

In time, a better description of the itinerary will be obtained :

* as more experience is gained with the conformity parameters.

* as other unit-level datasets provide new description points.

II.1.b.2. Working bottom-up

Another approach should be considered, which starts from the bottom (i.e. the unit-level data) and works up from there.

As no preliminary literature study is made, the itinerary description and the conformity rules will have to be distilled from the unit-level data. The layout of these rules will have been determined by the top-down prototype and will improve with time.

Thus, the bottom-up approach can be used only after the top-down work has set up and validated the model and extensive tests have been run with unit-level data. This follows the standard modelling procedure :

* start with well-known data but pretend you have no knowledge about them

* develop a model and check if this model matches the real-life data

* test the model on wider (but still well-known) datasets for improving it (learning dataset)

* use the model with some confidence on unknown data

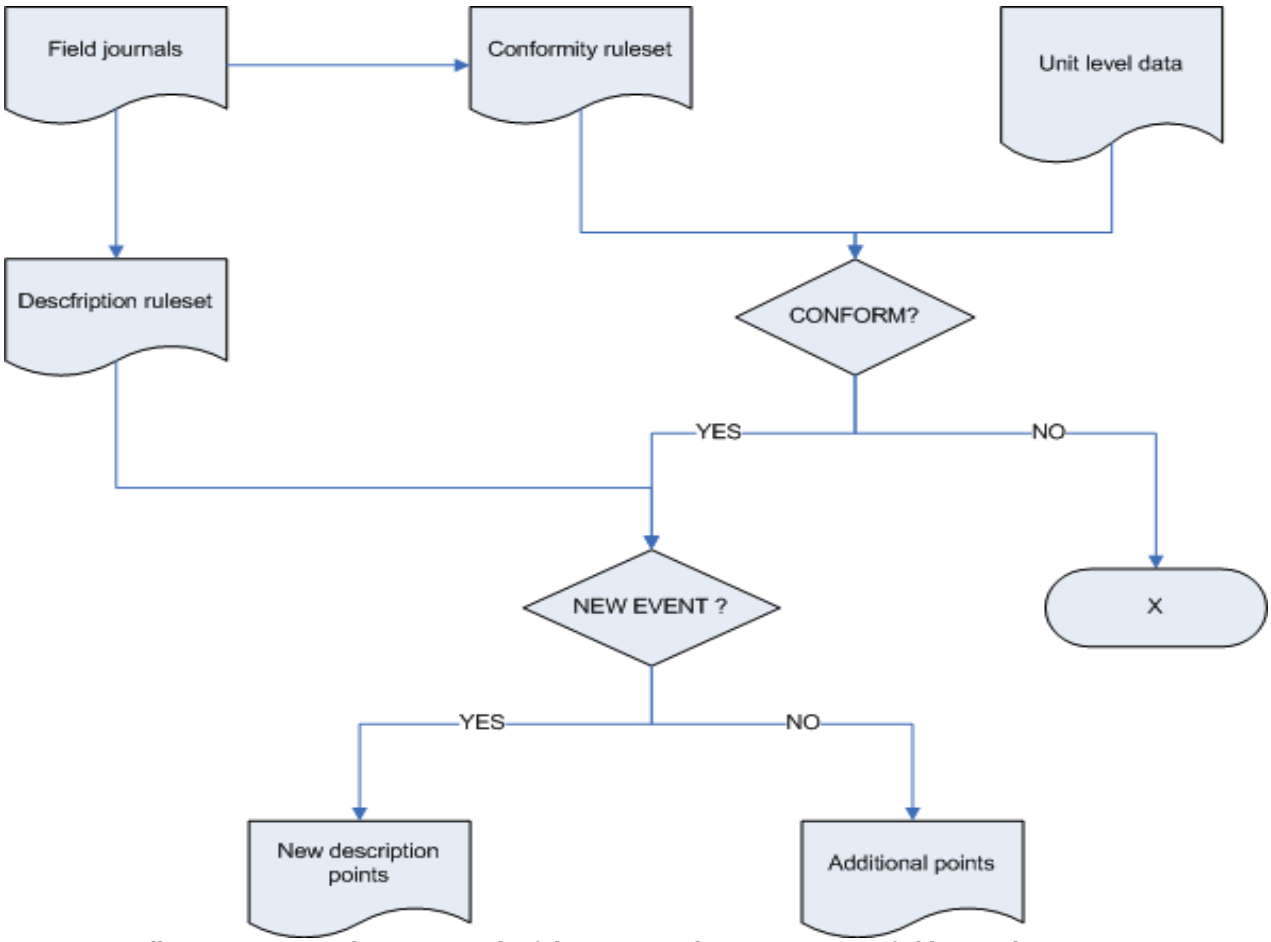


Illustration 3: Top-down approach of the itinerary description. From field journal information is made a conformity ruleset and an itinarary description ruleset. The unit level data are matched with the conformity ruleset to check if they could be part of the itinerary.If so, they are matched with the description ruleset to see if they add substantial extra information, and should be added to the itinerary's description points.

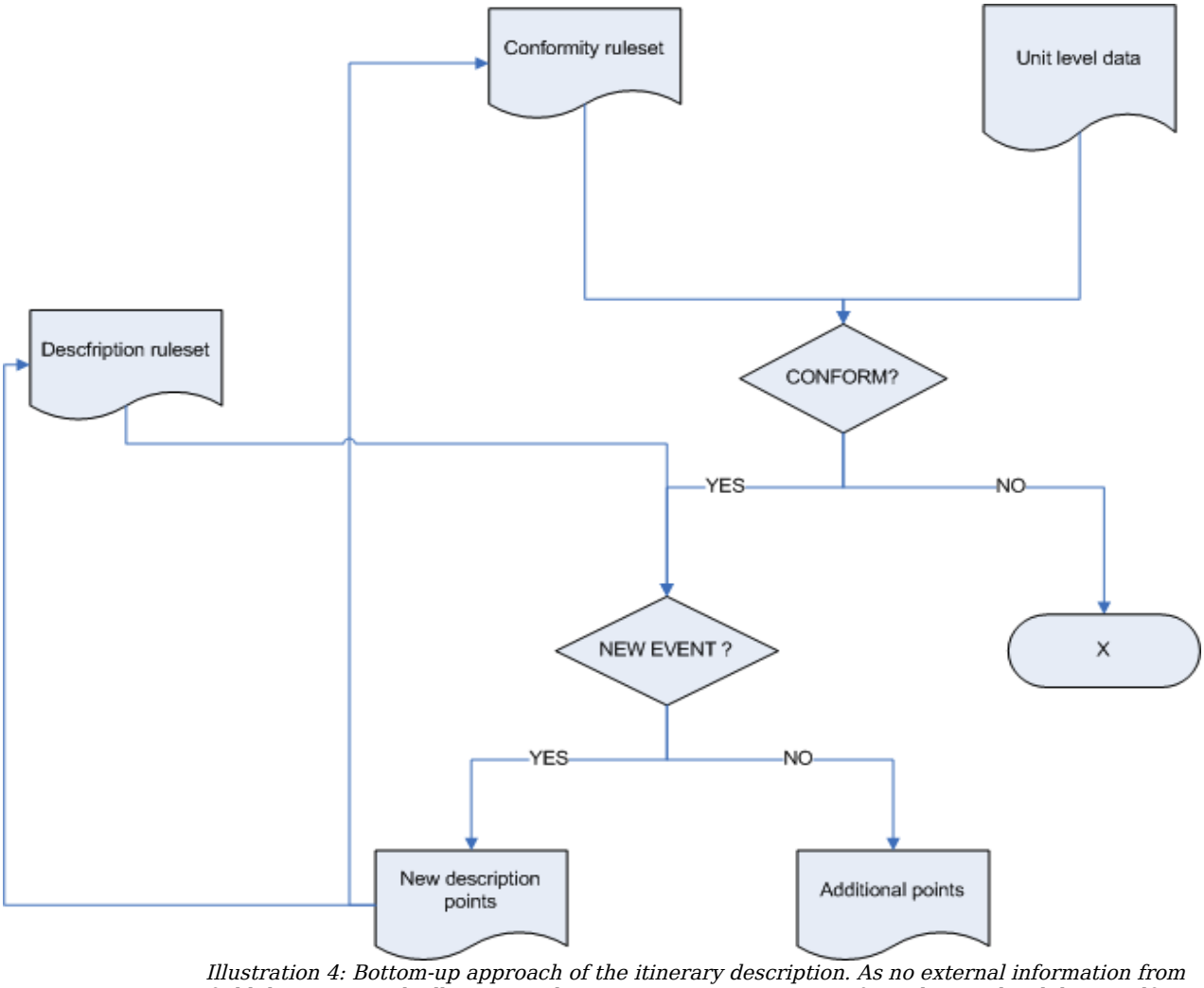


Illustration 4: Bottom-up approach of the itinerary description. As no external information from field diaries is used, all itinerary description points must come from the unit level data itself. A small number of initial conformity and description rules filters out data points which are conform, and new events (new description points). With each additional run, a more accurate description of the itinerary is obtained.

II.2. Analytic methodology

II.2.a. A formal description of itineraries in UML.

In order to get a clear overview of the subject of our study, a formal definition of "itineraries" and all related concepts (expeditions, participants, events,...) is needed.

For this task, the Unified Modelling Language (UML) was chosen :

* it offers a standard graphical notation for constructing an object model.
* it is possible to convert UML to and from Geographic Markup Language, GML . See Suzuki & Yamamoto, 1998, Portele, 2005 (1) and (2), Grønmo *et al.*, 2002).
* the UML schema comes very close to a programming pseudo code, assuring a straightforward practical implementation (Cox *et al.*,2002, Cox *et al.*, 2003, Martin,1998)

II.2.a.1. Expedition.

An expedition is an organised voyage or journey undertaken by a group of people with a scientific purpose.

attributes of an expedition :

* begin and end date
* publication(s) used to obtain this info
* expedition names (with an indication if a particular name is preferred for display)
* regions where the expedition has been
* participants

II.2.a.2. Participant

A participant is someone who takes an active part in an expedition or event.

attributes of a participant :

* last name and first name
* initials
* title
* address
* institution
* date of birth/death

II.2.a.3. Expedition event

An expedition event is something that impacts an expedition.

attributes of an expedition event :

* an event number
* a reference to the locality where the event happened
* a date and time for the start of the event
* a date and time for the end of the event

Some examples of expedition events :

* an arrival in a village or city

* passing a special landmark (place)* a change of transport (e.g. getting into a boat)

* spending the night, making camp

An expedition_event is georeferenced. This can be done to a single latitude / longitude with a measurement of uncertainty. A possible method would be the point-radius method as described by Wieczorek *et al.*, 2004. As this method is also used by the HerpNet network, to which the RMCA is participating, it will be tested on our problematic. More complex georeference methods, e.g. to describing polygons, can be used as well.

An expedition event is attached to zero or one sections (see below) and is subject to event constraints (see below).

II.2.a.4. Event_constraints

All information about constraints placed on events.

Constraints can be imposed by other events, by known dates, or by other reasons.

some examples of event constraints :

* you can't leave a village before you arrived there

* you can't use a bridge if it hasn't yet been built

* an event in "July" will have happened between July,1st and July, 31st

II.2.a.5. Event_constraint_type

Constraints come in different types, the most important being :

“not after " (date/event)

“not before" (date/event)

"span"

"window"

Were "span" means a period covering the full extent from end to end (staying at a village for two weeks, descending a river by boat for many days) and "window" covering the possible period within which the event occurred (e.g. the party passes the equator somewhere within that week).

II.2.a.6. Section

A section is a link between two consecutive expedition events, with the same participants but with different localities. A section is linked to just one itinerary.

II.2.a.7. Itinerary branching

An itinerary branching is a link between two consecutive expedition events, with the same locality but with different participants. An itinerary_branching is never linked to an itinerary. Instead, it marks the point in place and time where one itinerary stops and another (or others) begins.

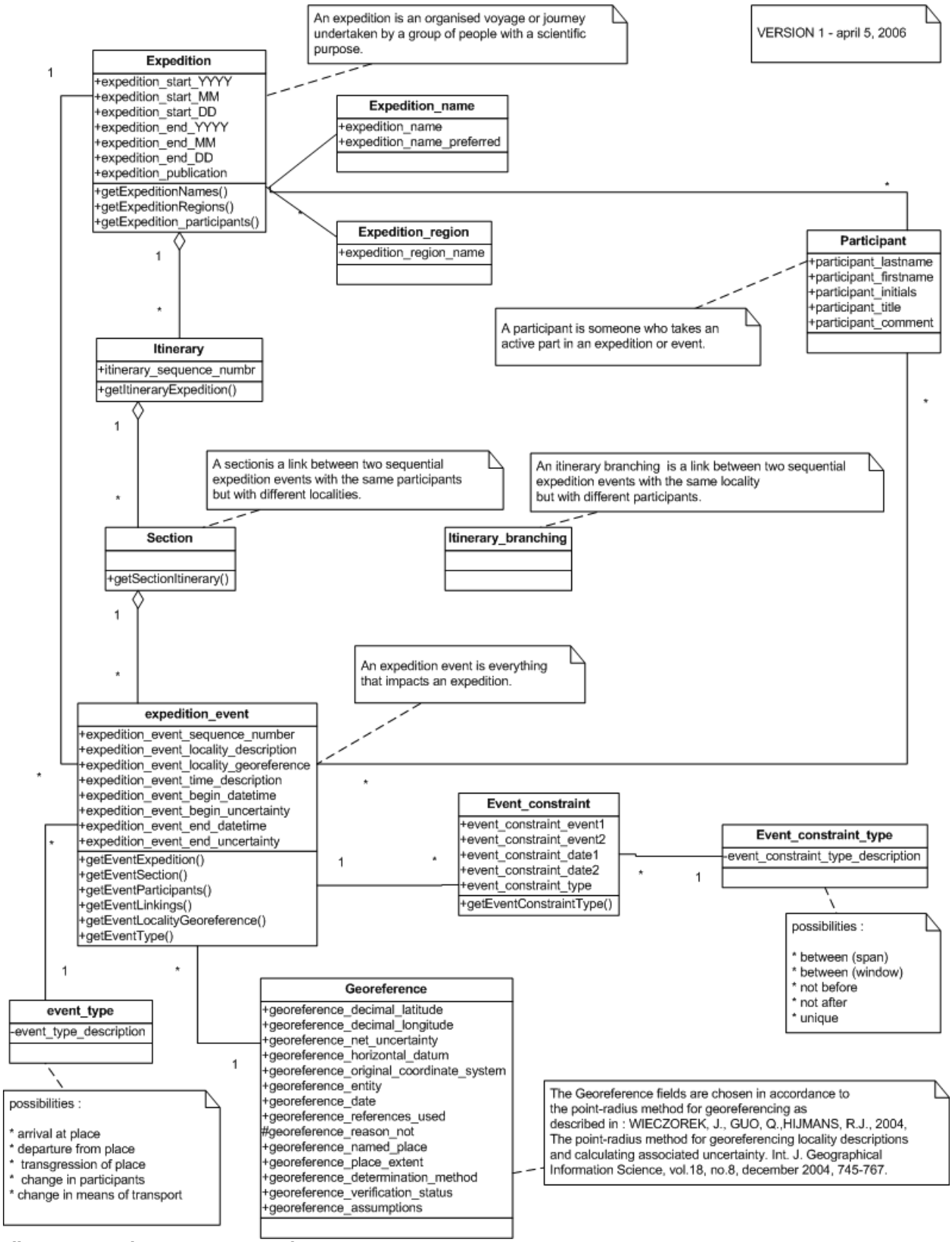


Illustration 5: The UML itinerary schema.

II.2.b. Identification of concerned ABCD 2.06 fields.

(see Berendsohn, 2005, and http://www.bgbm.org/TDWG/CODATA/Schema/)

As the itinerary project will preferably use ABCD as a data exchange standard, an identification has been made of which fields are relevant to the project. "Itineraries" is a new concept, but in diverse ABCD fields, relevant information is yet provided The table gives an overview of usable fields.

ABCD concepts	could be used for
Dataset/Metadata/Description/Title	expedition name
Dataset/Metadata/Description/Details	expedition description, possibly name
Dataset/Metadata/Description/Coverage	expedition region
Dataset/Metadata/Scope/GeoecologicalTerms	expedition region
Unit/NamedCollectionsOrSurveys	expedition name
Unit/Gathering/Datetime	event datetime / startdate / end date
Unit/Gathering/Agents	(gathering) event participants
Unit/Gathering/Localitytext	(gathering) event locality
Unit/Gathering/NamedAreas	(gathering) event locality
Unit/Gathering/NearNamedPlaces	(gathering) event constraint
Unit/Gathering/GML	visualisation
Unit/Gathering/WFS	visualisation and data sharing in OGC webservices
Unit/Gathering/SiteCoordinates	(gathering) event georeferencing
Unit/gathering/Altitude	(gathering) event georeferencing
Unit/Gathering/CollectorsFieldNumber	conformity rule set

However, not all concepts of the itinerary project are available in a usable form.In particular, the central notion of "event" seems to fit only partially in the ABCD concept of "Gathering". An event as specified in the itineraries can encompass a broad range of possibilities, of which a gathering event is only one (non exhaustive list):

* gathering events
* setting up camp
* changing of transport
*changing of participants

Perhaps the ABCD concept of "Gathering" could be embedded in a broader notion to cater for this.

The "event_type" and "event_constraint" concepts are specifically constructed to provide extra means for pinpointing and describing an event (gathering events and others). However, if an extension of the "gathering" concept into a broader "event" notion would be considered, "event-type" and "event-constraint" could be a valuable addition to the schema, in describing itineraries and undoubtedly in many other occasions. For example the information now contained in "Unit/Gathering/NearNamedPlaces" (describing a constraint

on the gathering location), could be relocated here.

"Itinerary", "section" and "itinerary branching" are even more specific, and currently seem to have no correspondent concepts in the ABCD scheme. They could be implemented as additional concepts, with their appropriate links to the enlarged "event" element.

"Expedition" and "participant" notions seem adequately provided in the schema as it is today.

II.2.c. Choosing a testcase : the Lang and Chapin journey.

(After Slack, G., see http://diglib1.amnh.org/intro/intro.html)

For testing assumptions and techniques, a well-known expedition was selected to serve as a typical example : the Lang and Chapin expedition to (then) the Belgian Congo, 1909-1915. This was the first comprehensive survey of North-eastern Congo, initiated by the American Museum for Natural History (AMNH) in cooperation with the Belgian authorities. It sought to capture as broad as possible a picture of the Congo's biota and cultures, in particular of the rare and at that time considered strange okapi (*Okapia johnstoni (*Sclater*,* 1901*)*). In five years' time, it brought home some 5,800 mammals, 6,400 birds, 4,800 reptiles and amphibians, 6,000 fish, over 100,000 invertebrates, and 3,800 anthropological objects. Duplicate specimens were given to the Royal Congo Museum (now RMCA). The AMNH website presents an interesting visualisation of this journey, as well as additional information.

It was an interesting choice for a testcase :

* extensive expedition diaries are available

* duplicate specimens are in the collection of the RMCA

* the pathway of the expedition is well-known

* the journey has interesting features such as expedition splits and joins

II.2.d. Technical methodology

II.2.e. Technologies and formats

In accordance with TDWG and BioCASE philosophies, open standards and open software have been chosen and used preferably for the itinerary project. The reports mentioned earlier (Beller *et al.*, 2004. Krumenacker & Malicky, 2004, Torre, 2005) took a closer look at some possible tools, and made recommendations that were very helpful in the decision process

II.2.f. Data standards.

II.2.f.1. ABCD

The Access to Biological Collections Data (ABCD) Schema is an evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data). (http://www.bgbm.org/TDWG/CODATA/Schema/))

II.2.f.2. XML

XML is a World Wide Web Consortium (W3C) standard for data exchange. For clarity and standardisation, we prefer to use an XML-application for all files used, where possible.

II.2.f.3. GML

The Geography Markup Language (GML) is an XML encoding for the modelling, transport, and storage of geographic information including both the spatial and non-spatial properties of geographic features (Krumenacker & Malicky, 2004).

* The DeeGree WMS (see below) is fully compliant with GML
* PostGIS (see below) has support for GML loading and export.
* GML can be generated from UML (see above).

II.2.g. Tools.

II.2.g.1. PostGreSQL

A robust SQL database was needed for storing and manipulating the various testing datasets in a standardised manner.

* free & open source
* can run PostGIS, so the link to the DeeGree server can be made.
* fast & reliable, can be scaled up for serving over the Net.
* excellent graphical management tools available (PgAdmin).
* exists for multiple platforms

II.2.g.2. PostGIS

PostGIS provides a set of geographic extensions to the PostGreSQL database, to spatially enable it (http://www.postgis.org/). Thus, manipulation of geospatial features becomes possible. A PostGIS-enabled database can feed data to the DeeGree server.

* free & open source

* can connect to the DeeGree server, so make the link from database to WMS.

II.2.g.3. DeeGree

See also Fitzke *et al.*, 2003.

Deegree is a Java framework offering the main building blocks for spatial data infrastructures. Its entire architecture is developed using standards of the Open Geospatial Consortium (OGC) and ISO/TC 211. Deegree encompasses OGC Web Services (http://www.deegree.org/).

DeeGree was prioritised by the ENBI Report on proposed data standards and protocols with respect to analysis tools (WP10-D10.1b) (Krumenacker & Malicky, 2004).

* free & open source

* runs on all platforms

* iGeoportal can be installed as a graphical user interface

II.2.g.4. Apache Tomcat

Apache Tomcat is the servlet container that is used in the official Reference Implementation for the Java Servlet and JavaServer Pages technologies (http://tomcat.apache.org/). Tomcat is required for handling the DeeGree servlets.

* free & open source

* runs on all platforms

* closely integrated with Apache web server

II.2.g.5. UMLEditor

UMLEditor is a Java tool for the construction of UML schemas.

* free and open source

* Java, runs on all operating systems

II.2.g.6. DivaGIS

DivaGIS is a free and open source GIS tool.

* free and open source

* currently Windows-only, a Java version is under way

* contains already embedded applications for basic data quality control (i.e. checking if geographic coordinates are indeed within the boundaries of the database indicated named area).

II.2.g.7. QGIS

Quantum GIS (QGIS) is an open source GIS tool.

* free and open source

* runs on all platforms

* supports vector, raster, and database formats.

* very easy for rapidly importing a spreadsheet file (.csv) with points for visualising.

II.2.g.8. Google Earth

Google Earth is a commercial interactive service for exploring a globe where satellite and

aircraft images are mapped.A free version is available for home use, but extra features require a license fee. With a proprietary XML-like language (KML / KMZ), own georeferenced material can be displayed. An example is the very precise itinerary of the Iditarod Alaskan dog sled race, as presented on the EarthSLOT site.

A whole journey can be presented, and individual points can be selected for more details, and links to further information.

Illustration 6: Itinerary of the Iditarod dog sled race. Mapping to Google Earth by Peter Prokein, University of Alaska Fairbanks.

Illustration 7: The Iditarod dog sled race. Selection of a point to show its details. Mapping to Google Earth by Peter Prokein, University of Alaska Fairbanks.

NASA Worldwind

NASA Worldwind is a world visualisation tool very similar to Google Earth.

However, it has a different philosophy : it is free, open source, and more scientifically oriented than Google Earth. Also, it is only available for the Windows platform as of this writing. For a comparison between Google Earth and Worldwind, see http://www.worldwindcentral.com/wiki/Google_Earth_comparison

A good example of the use of Worldwind is the "Darwin" add-on, which maps the famous voyage of the H.M.S. Beagle carrying Charles Darwin.

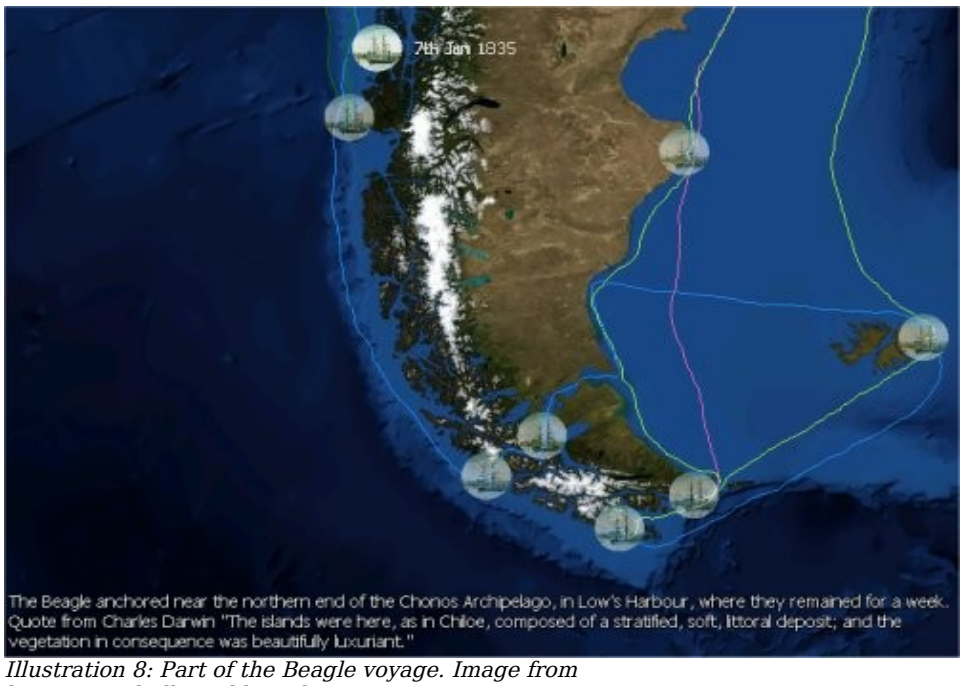


Illustration 8: Part of the Beagle voyage. Image from http://www.bullsworld.co.uk/

III. Results

III.1. An algorithm for aggregating unit-level data : a prototype

Preliminary note :

For all formulae given underneath the definition of all parameters, variables and fixed values is as follows :

* x, y : latitude and longitude coordinate of an abstract point P
* u(s)P : spatial uncertainty for point P
* x1, y1 : latitude and longitude coordinate of a specific point P1
* u(s)P1 : spatial uncertainty for point P1
* x2, y2 : latitude and longitude coordinate of a specific point P2
* u(s)P2 : spatial uncertainty for point P2
* Max_Temporal_Uncertainty : parameter for setting a maximal allowed temporal uncertainty on a per-point basis.
* Distance_per_Day : calculated distance travelled per day by the expedition party.
* Average_Distance_per_Day : a parameter for setting the average distance assumed to be possibly travelled in a day.
* Max_Distance_per_Day : parameter for setting a maximal allowed distance per day, for filtering out "impossible" distances, or detecting changes of transport.
* Total_Distance : grand total of calculated travelled distances
* Total_Number_of_Points : grand total of description points used
* Total_Spatial_Uncertainty. The simple um of all uncertainties of the points = sum(u(s)P1...u(s)Pn)

III.1.a. The conformity rule set

As explained, the conformity rule set is used to decide which records possibly belong to an expedition. This is a boolean yes-or-no decision.

The data are sorted on date. All comparisons are made in relation to the previous point, the very first point being excluded from the demand to be conform.

Other records are considered to be conform if the following conditions are met :

Time and date are sufficiently precise :

u(s)P < Max_Temporal_Uncertainty

The initial value of Max_Temporal_Uncertainty will be 1 day, thus keeping the most accurate points, and presumably most of the camp-making events. Note that a time span does not count as "uncertainty".

The speed since the previous point doesn't exceed the fixed limit for possible distance per day

Distance_per_Day < Max_Distance_per_Day

An initial value for Max_Distance_per_Day would be 50 km, assuming journeys on foot.

If the distance to the previous point exceeds the limit, several actions can be taken :

* a warning is issued : human intervention is then needed to decide. In particular, a first run will flag some "impossible" day distances which can be easily attributed to fast means of transport (e.g. train, boat,...)
* an automatic "break", and start of a new itinerary. This caters for the situation where some links in the journey are not known. The dataset is then effectively split in two.
* The point is discarded altogether.

The total (combined) spatial uncertainty of the end- and beginning point is less than the distance travelled

Distance_Travelled < u(s)P1 + u(s)P1

These pathways are too unsure to be relevant.

III.1.b. Conformity score :

Apart from the decision conform/not conform, a score is given to decide how well the conform points fit together. This score is then used for constructing the most probable pathway. For this score, four parameters are considered :

The total distance travelled

Total_Distance

A total journey length that is too long may point to improbable constructions, or erroneous data points.

The number of points :

Total_Number_of_Points

More points describing the same distance increase the accuracy.

The total uncertainty of the points :

Total_Spatial_Uncertainty

Where the uncertainty values of the middle points (i.e. all but the first and last point) are counted double, because they work in two directions.

The total "slack"

"Slack" gives a measure for the maximum possible deviation of the straight path between two points. It compares the calculated distance between the two points with the theoretical distance that could have been covered in the given time (calculated from the Ag_Speed_per_Day parameter). The surplus gives a degree of freedom for all sorts of deviations and detours between the two points. Thus, more "slack" means the actual pathway is less well known.

Total_Slack = Average_Distance_per_Day - Distance_per_Day

These four parameters can be weighed to become a single conformity score :

Conformity Score (CS) = (Total_Distance + Total_uncertainty / Total_slack) / (Number_of_Points-1)

Although all relevant parameters are included in this score, no physical meaning should be given to it, nor should a unit be added. It is not more than an indicator for making

comparisons.

The CS value can be calculated for any number of points, with a lower score indicating a better 'fit' of the points.

III.1.c. Examples of the Conformity Score.

Some examples will present the idea of the Conformity Score, and how it changes with various situations.

The Conformity Score should of course ideally be calculated with the real distances between the points, all refering to the same datum (i.e. WGS84), with the appropriate algorithms.

In the present preliminary check, however, as the latitudes are all within a few degrees from the equator, a simplified distance calculation by Euclidian measurement was adopted :

distance = SQRT((lat1+lat2)² + (lon1+lon2)²)

For latitudes less than 30 degrees, and distances of about 20 km, this simplification has an error of less than 9 meters.

Further improvements could use spherical distances : the Haversine formula (from Sinnott, 1984) for a spherical Earth with radius R, two points (lon1, lat1) en (lon2, lat2).

Calculate their differences in latitudes and longitudes :

dlon = lon2 – lon1 (in radials)

dlat = lat2 – lat1 (in radials)

then calculate their distance :

distance = R * 2 * arcsin(min[1,sqrt(a)])

where a= sin^2(dlat/2) + cos(lat1)*cos(lat2)*sin^2(dlon/2)

and R can be approximated by 6378 - 21*sin(lat) km.

for lat can be used : (lat2-lat1)/2.

And better still, take into account the datum. According to Snyder, 1987 a good approximation using the oblate spheroid as defined by e.g. WGS84 would replace R by R'

R'= a * (1-e²)/(1-e² * sin²(lat)) ^ (3/2)

where a=equatorial radius, b=polar radius and e = (1 – b²/a²)^(1/2)

but the use of R' = a – (b-a) * sin(lat) proves to be mostly accurate enough

for lat can be used : (lat2-lat1)/2.

The use of a simplified dinstance calculation in the eventual algorithm (Euclidian or spherical distance, use of WGS84 or simple lat/long) will be evaluated by their impact on the Conformity Score and the calculation time needed in realtime conditions : strongly iterative approaches are naturally not indicated.

III.1.c.1. Example 1: An actual expedition itinerary

locality	longitude	latitude	i (deegrees)	i (km)	date begin	date end	distance (km)	traveling days	km/day
Bafwaboli	26.17	0.65	0.02	2.22	09/12/09	09/12/09			
Bafwasende	27.27	1.08	0.02	2.22	09/24/09	09/24/09	131.23	12	11
Avakubi	27.57	1.33	0.02	2.22	09/30/09	12/07/09	43.35	6	7
Ngayu	27.55	1.75	0.02	2.22	12/10/09	12/26/09	46.29	3	15
Medje	27.3	2.42	0.02	2.22	01/13/10	10/15/10	79.03	18	4
Pawa	27.7	2.53	0.02	2.22	10/18/10	10/18/10	46.15	3	15
Isiro	27.68	2.8	0.02	2.22	10/23/10	10/23/10	30.03	5	6
Nala	27.67	2.87	0.02	2.22	10/26/10	10/26/10	7.63	3	3
Rungu	27.88	3.18	0.02	2.22	10/28/10	10/28/10	42.59	2	21
Niangara	27.88	3.7	0.02	2.22	11/01/10	01/20/11	57.35	4	14
Dungu	28.57	3.62	0.02	2.22	01/25/11	01/30/11	76.41	5	15
Faradje	29.7	3.75	0.02	2.22	02/06/11	02/19/13	126.67	7	18

Table 1: A part of the Lang & Chapin journey in the Kisangani (Stanleyville) region

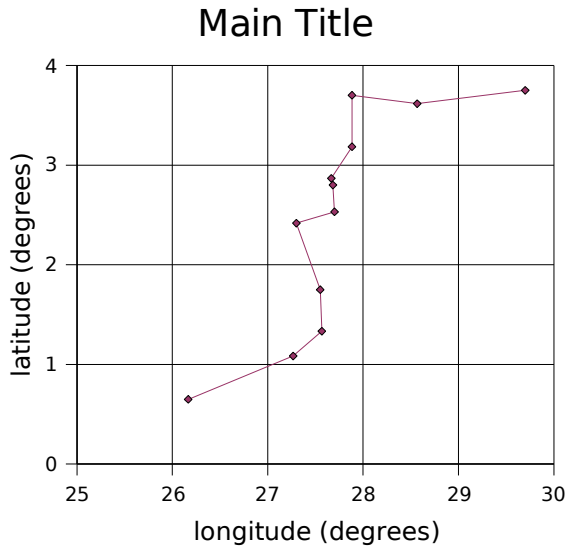


Illustration 9: Geographical plot of the points from the table. Bottom left is Bafwaboli, top right is Faradje.

distance :	686.72
points-1 :	11
km/point :	62.43
inacuracy:	48.84
slack :	1353.28
CS :	0.049

A part of the Lang & Chapin expedition. Latitudes and longitudes are taken from the field journals. A standard accuracy of 0.02 decimal degrees has been assigned, roughly 2 kilometres The graph shows a 700 km journey, from Bafwaboli to Faradje.

III.1.c.2. Example 2: addition of a conform point

locality	longitude	latitude	i (deegrees)	i (km)	date begin	date end	distance (km)	traveling days	km/day
Bafwaboli	26.17	0.65	0.02	2.22	09/12/09	09/12/09			
Bafwasende	27.27	1.08	0.02	2.22	09/24/09	09/24/09	131.23	12	11
Avakubi	27.57	1.33	0.02	2.22	09/30/09	12/07/09	43.35	6	7
Ngayu	27.55	1.75	0.02	2.22	12/10/09	12/26/09	46.29	3	15
Medje	27.3	2.42	0.02	2.22	01/13/10	10/15/10	79.03	18	4
Pawa	27.7	2.53	0.02	2.22	10/18/10	10/18/10	46.15	3	15
Isiro	27.68	2.8	0.02	2.22	10/23/10	10/23/10	30.03	5	6
Nala	27.67	2.87	0.02	2.22	10/26/10	10/26/10	7.63	3	3
Rungu	27.88	3.18	0.02	2.22	10/28/10	10/28/10	42.59	2	21
Niangara	27.88	3.7	0.02	2.22	11/01/10	01/20/11	57.35	4	14
Dungu	28.57	3.62	0.02	2.22	01/25/11	01/30/11	76.41	5	15
extra point	**29**	**3.5**	**0.02**	**2.22**	**02/03/11**	**02/03/11**	**49.81**	**4**	**12**
Faradje	29.7	3.75	0.02	2.22	02/06/11	02/19/13	82.51	3	28

Table 2: A part of the Lang & Chapin journey in the Kisangani (Stanleyville) region, with an extra added point at (3.5, 29).

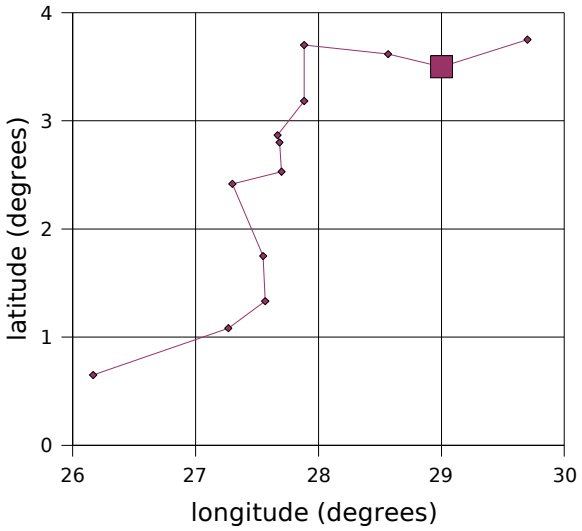


Illustration 10: Geographical plot of the points from the table. Bottom left is Bafwaboli, top right is Faradje.

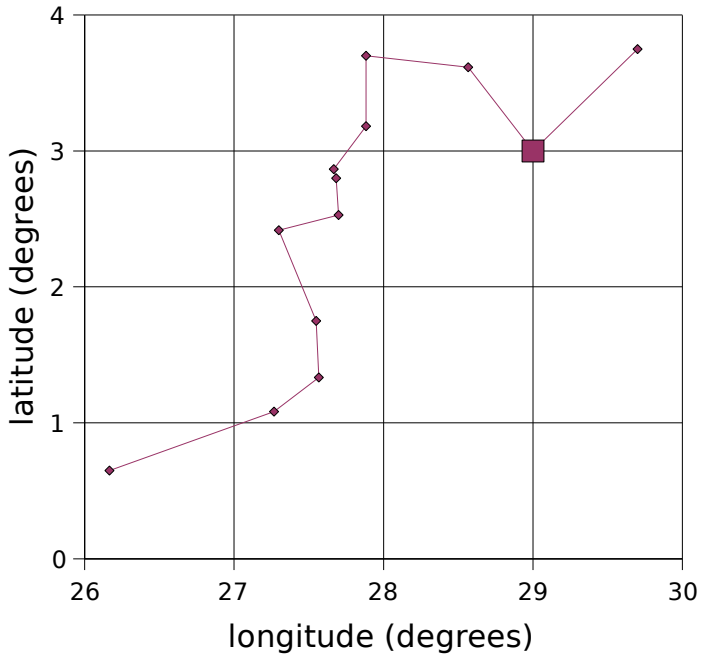
distance :	692.37
points-1 :	12
km/point :	57.7
inacuracy:	53.28
slack :	1347.63
CS :	0.046

The same journey with an extra added (fictitious) point at (3.5,29). The total length of the journey is slightly increased (some 6 km), but the CS score is better due to the additional information the point provides (0,046 versus 0.049)

III.1.c.3. Example 3 : addition of a less conform point

locality	latitude	longitude	i (deegrees)	i (km)	date begin	date end	distance (km)	traveling days	km/day
Bafwaboli	0.65	26.17	0.02	2.22	09/12/09	09/12/09			
Bafwasende	1.08	27.27	0.02	2.22	09/24/09	09/24/09	131.23	12	11
Avakubi	1.33	27.57	0.02	2.22	09/30/09	12/07/09	43.35	6	7
Ngayu	1.75	27.55	0.02	2.22	12/10/09	12/26/09	46.29	3	15
Medje	2.42	27.3	0.02	2.22	01/13/10	10/15/10	79.03	18	4
Pawa	2.53	27.7	0.02	2.22	10/18/10	10/18/10	46.15	3	15
Isiro	2.8	27.68	0.02	2.22	10/23/10	10/23/10	30.03	5	6
Nala	2.87	27.67	0.02	2.22	10/26/10	10/26/10	7.63	3	3
Rungu	3.18	27.88	0.02	2.22	10/28/10	10/28/10	42.59	2	21
Niangara	3.7	27.88	0.02	2.22	11/01/10	01/20/11	57.35	4	14
Dungu	3.62	28.57	0.02	2.22	01/25/11	01/30/11	76.41	5	15
extra point	**3**	29	0.02	2.22	02/03/11	02/03/11	83.66	3	28
Faradje	3.75	29.7	0.02	2.22	02/06/11	02/19/13	113.88	4	28

Table 3: A part of the Lang & Chapin journey in the Kisangani (Stanleyville) region, with an extra added point at (3, 29)



distance :	757.59
points-1 :	12
km/point :	63.13
inacuracy:	53.28
slack :	1282.41
CS :	0.053

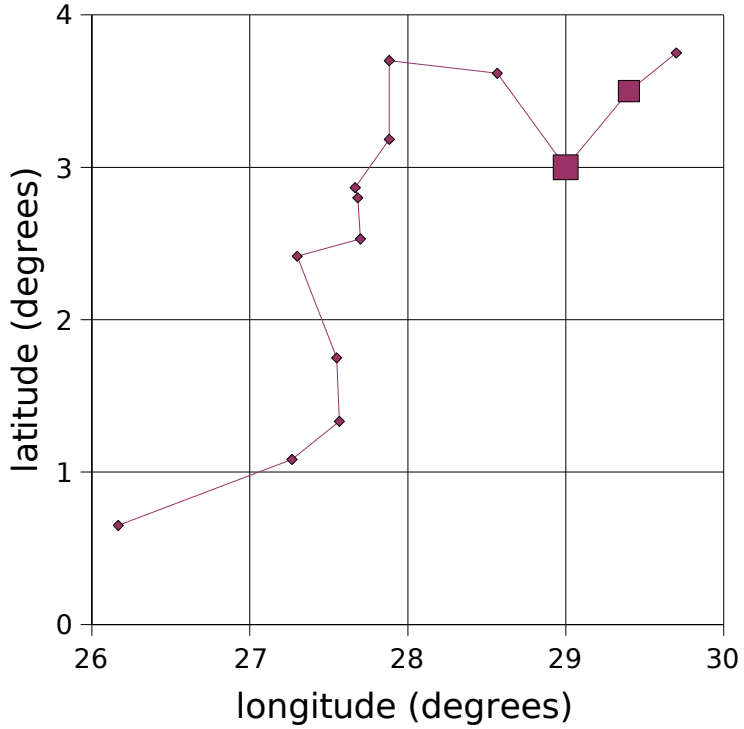
Illustration 11: Geographical plot of the points from the table. Bottom left is Bafwaboli, top right is Faradje.

The same journey again, with an extra point at (3, 29), making a substantial detour.(some 75 extra kilometres).from the original situation. The CS is consequently bigger (0.053) in comparison to the starting situation (0.049).

III.1.c.4. Example 4 : addition of two points

locality	latitude	longitude	i (deegrees)	i (km)	date begin	date end	distance (km)	traveling days	km/day
Bafwaboli	0.65	26.17	0.02	2.22	09/12/09	09/12/09			
Bafwasende	1.08	27.27	0.02	2.22	09/24/09	09/24/09	131.23	12	11
Avakubi	1.33	27.57	0.02	2.22	09/30/09	12/07/09	43.35	6	7
Ngayu	1.75	27.55	0.02	2.22	12/10/09	12/26/09	46.29	3	15
Medje	2.42	27.3	0.02	2.22	01/13/10	10/15/10	79.03	18	4
Pawa	2.53	27.7	0.02	2.22	10/18/10	10/18/10	46.15	3	15
Isiro	2.8	27.68	0.02	2.22	10/23/10	10/23/10	30.03	5	6
Nala	2.87	27.67	0.02	2.22	10/26/10	10/26/10	7.63	3	3
Rungu	3.18	27.88	0.02	2.22	10/28/10	10/28/10	42.59	2	21
Niangara	3.7	27.88	0.02	2.22	11/01/10	01/20/11	57.35	4	14
Dungu	3.62	28.57	0.02	2.22	01/25/11	01/30/11	76.41	5	15
extra point 1	**3**	29	0.02	2.22	02/03/11	02/03/11	83.66	4	21
extra point 2	**3.5**	29.4	0.02	2.22	02/05/11	02/05/11	71.07	2	36
Faradje	3.75	29.7	0.02	2.22	02/06/11	02/19/13	43.35	1	43

Table 4: A part of the Lang & Chapin journey in the Kisangani (Stanleyville) region, with two



distance :	758.13
points-1 :	13
km/point :	58.32
inacuracy:	59.94
slack :	1281.87
CS :	0.049

Illustration 12: Geographical plot of the points from the table. Bottom left is Bafwaboli, top right is Faradje.

The same journey, with two additional points, at (3, 29) and (3.5, 29.4).The second additional point (3.5,29.4) supports the first one, lending it credibility as a valid journey point instead of an improbable detour. As a result, the CS is much lower (0.049 versus 0.053 for the single additional non-conform point (3,29)).

III.1.d. Discussion.

Though in the demonstrated cases the CS value seems to work fine, we should remark that it represents only a first step towards an encompassing algorithm for the itinerary decisions. Closer examination and testing exposed multiple drawbacks.Indeed, it could be argued that the algorithm in its present form is only useful in very specific cases. Its real value lies therefore in the lessons learned while working on it, and in careful analysis of its strong and weak points :

* In the present algorithm the sequence of points is determined only by their dates. Points occurring within a same day, or points without an exact date are therefore not considered. They can be integrated by adapting a temporal cumulative approach in a derived algorithm.

* the statistical behaviour of the CS value is yet unknown, and extensive testing might be necessary to obtain an idea of it. An alternative construction from a more statistical starting point would prove to be more appropriate. The search for relevant statistical methods will therefore be intensified.

* the algorithm uses absolutes and maximal values rather than probabilities (e.g. sum of squares). Therefore, its cutoff decisions are absolute (simple sums) and have not yet the gradation which they could have. This gradation will depend on the statistics that will be used.

* the basic premise is sound : a comparison between cumulative distances and expected distances to calculate probabilities With some minor changes it can be adapted for detecting the best possible connection of points within a same day.

During the practical implementation, with frequent test runs with the various kinds of data obtained, the algorithm will get the necessary improvement and sophistication. In that process, the above considerations and experiences will prove of great value.

III.2. Visualisations

III.2.a. Visualisations in QGIS

III.2.a.1. A part of the Lang and Chapin Belgian Congo expedition (1909-1915). Kisangani (Stanleyville) region

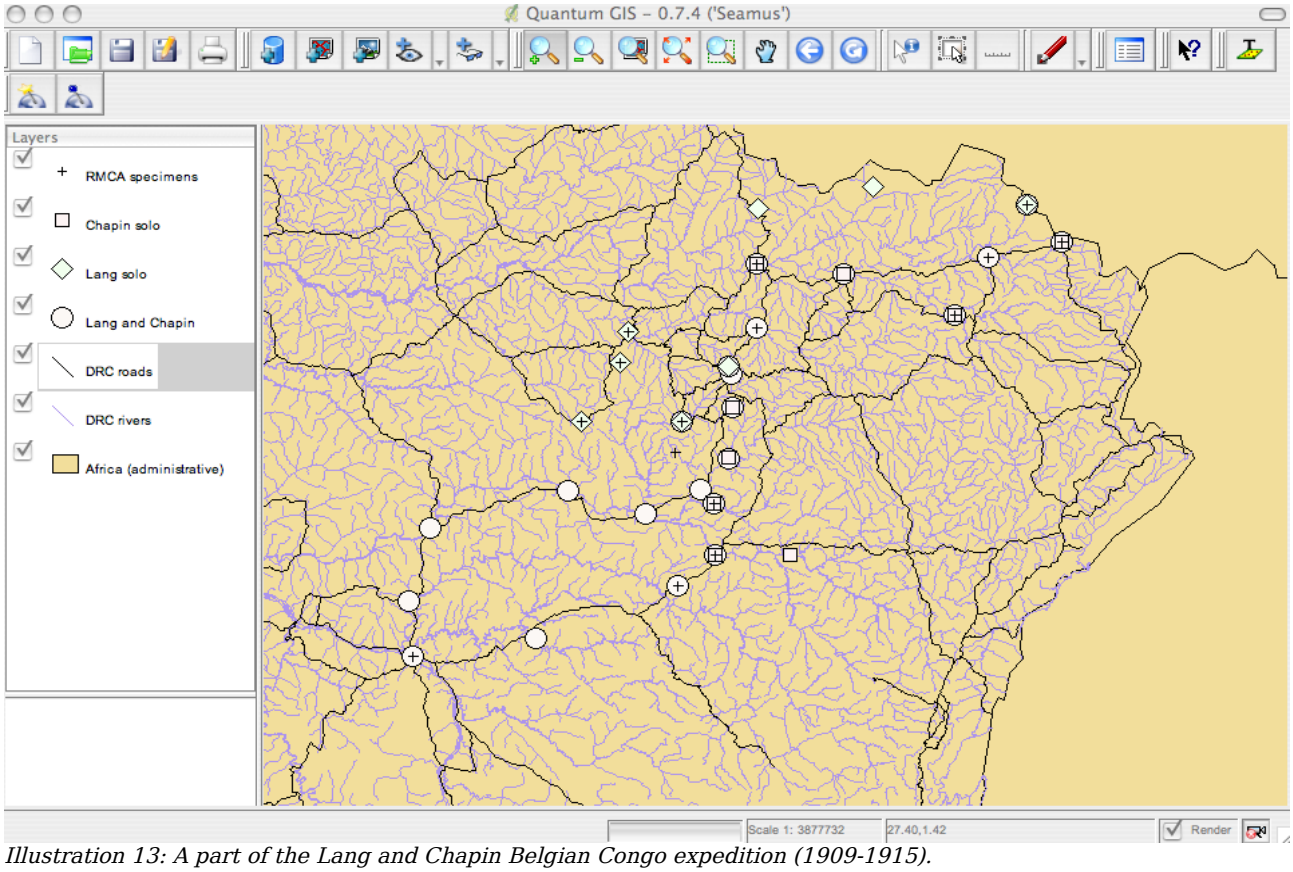


Illustration 13: A part of the Lang and Chapin Belgian Congo expedition (1909-1915).

Scale is 1/ 3.880.000 approx., middle of the map is near 27,7E 2N.

Map points from RMCA data and the AMNH website

Different points provide an easy overview of :

* places Lang and Chapin visited together.
* places visited by either Lang or Chapin on solo trips, after an expedition branching.
* places were RMCA specimens from that expedition have been collected.

One notices quickly :

* some of the RMCA "Lang and Chapin" specimens have in fact been gathered by Lang alone. Thus, our specimen label information could be annotated with additional information. The itinerary algorithms would report this possible refinement.

* some of the RMCA specimens have gathering locations seemingly unvisited by the Lang and Chapin expedition. So, further investigations have to be made : are all the lat/longs correct ? Were all the known expedition points included ? Perhaps the specimen was bought from somewhere else ?

III.2.a.2. A part of the Lang and Chapin Belgian Congo expedition (1909-1915) : Kinshasa (Leopoldville) region.

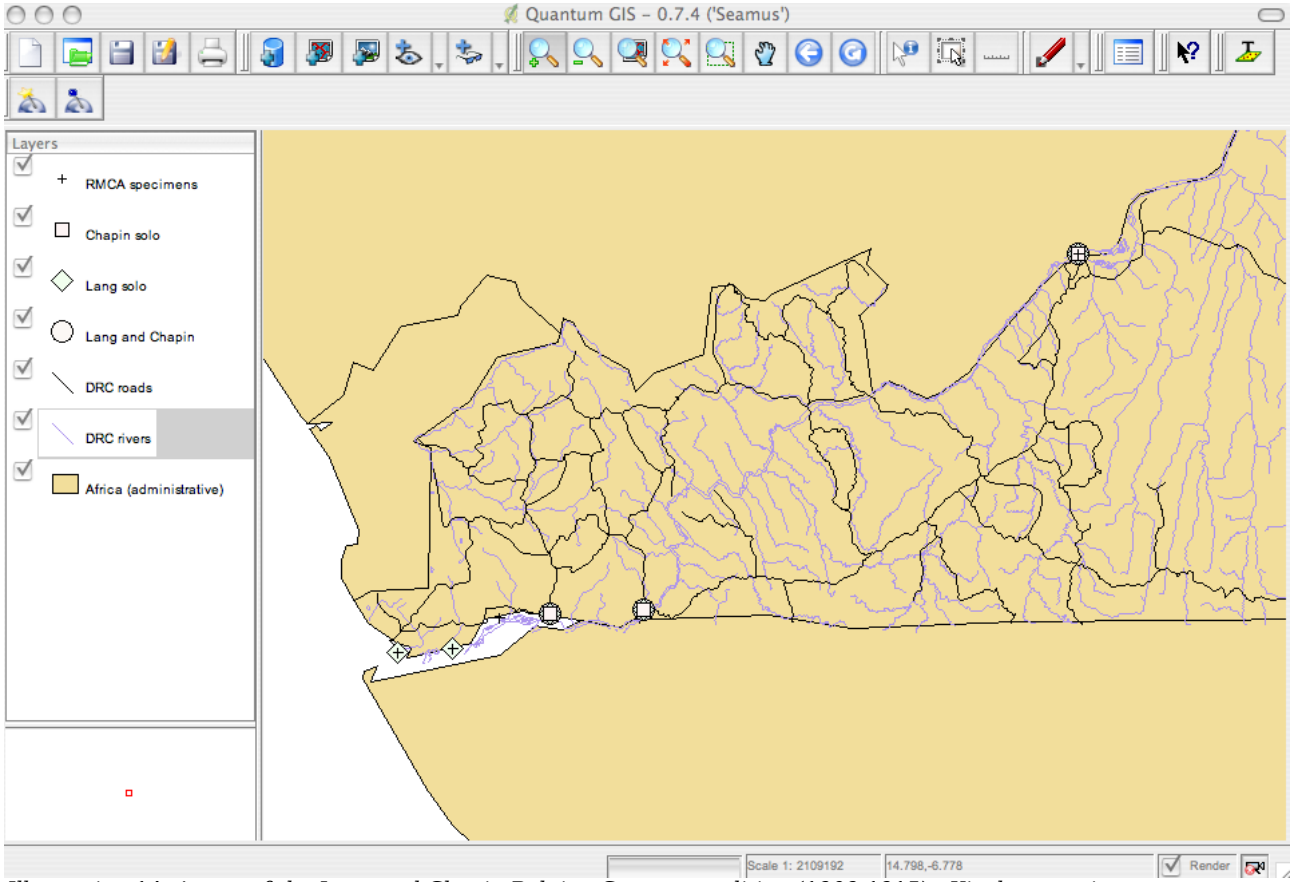


Illustration 14: A part of the Lang and Chapin Belgian Congo expedition (1909-1915) : Kinshasa region.

Scale is 1: 2.100.000 approx., middle of the map is near 14E 5S.

Map points from RMCA data and the AMNH website (ref)

Here, RMCA "Lang and Chapin" specimens have in fact been gathered by Lang alone, on his (solo) way back.

For example, the leftmost point (Banana) comprises 2 RMCA specimens currently labelled : "Lang and Chapin 7/1915". In the database, this information can be refined to : "Lang, 25-31/7/1915". An new label with an annotation could be added to the specimen.

The point just at the right (Boma) marks a RMCA specimen only dated to the year : "Lang and Chapin 1915". This can be refined to within a period of 10 days : "Lang, 2-12/7/1915"

The itinerary software will be able to suggest these refinements automatically

III.2.b. Visualisations in Deegree WMS and iGeoportal

III.2.b.1. Deegree Web Map Service : displaying maps

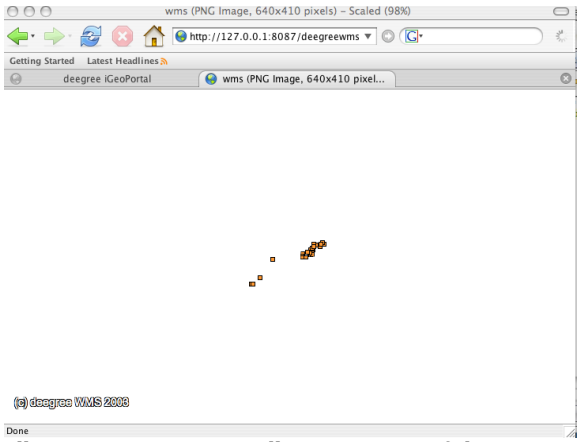


Illustration 16: Some collection points of the Lang and Chapin expedition.

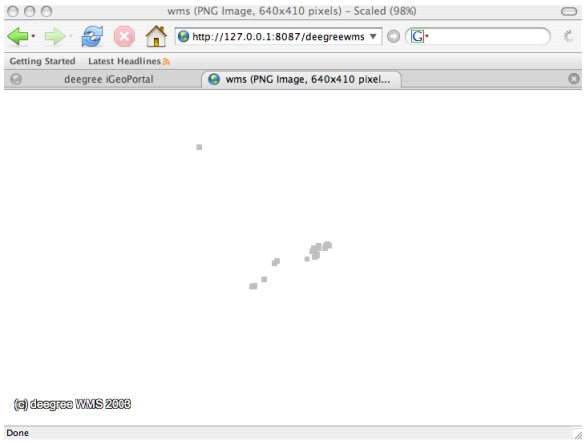


Illustration 15: Some collection points of the RMCA specimens from the Lang and Chapin expedition.

The DeeGree Web Map Service server enables the display of maps on the network. Adaptable parameters allow for requesting different parts of a map. Also, a map request can dynamically be transmitted to a database, and the map can be generated accordingly. So there is flexibility in the display. The pictures give three different maps, generated from QGIS, stored as shapefiles (.shp format) and served through DeeGree WMS.

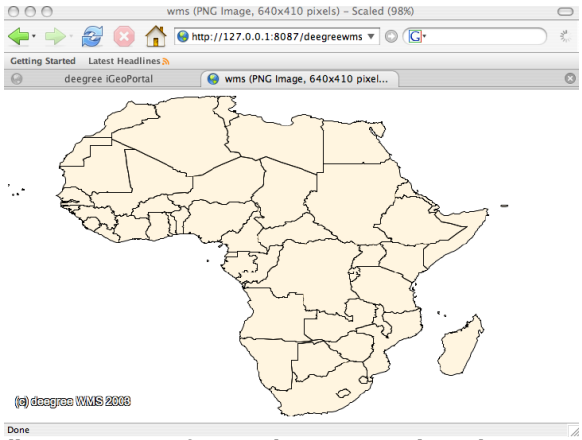


Illustration 17: Africa : administrative boundaries. Points taken from http://carpe.umd.edu/index.asp

III.2.b.2. iGeoportal : an interactive interface to the DeeGree WMS

iGeoportal provides an interactive graphical interface for accessing the DeeGree server (and other WMS-servers as well). For the itinerary project, it is a valuable addition, as it allows maps to be browsed and altered (scale, position), and individual layers to be queried.

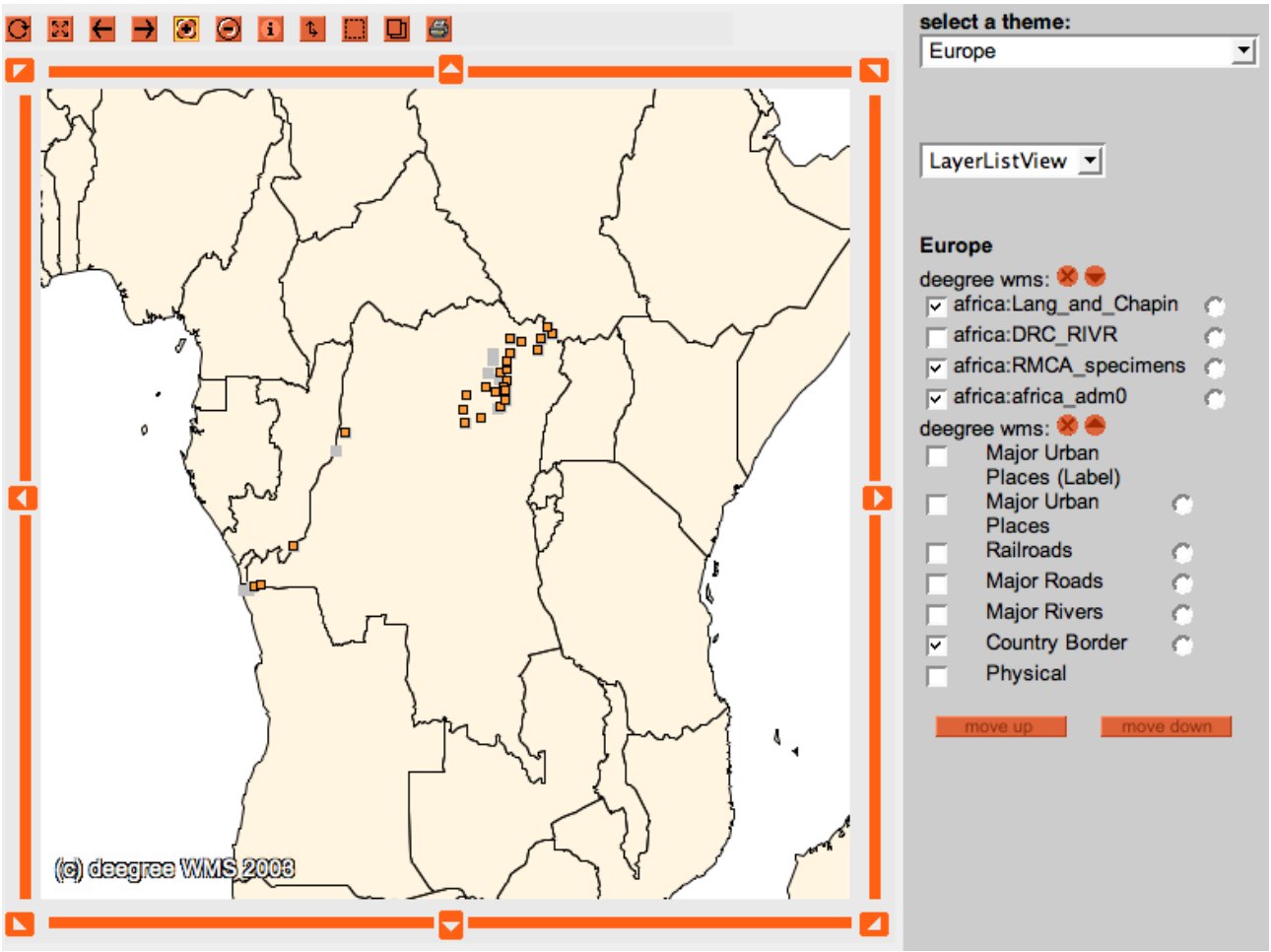


Illustration 18: A visualisation of 3 WMS-layers in iGeoportal. Middle of the map is near 24E,3S.

The screenshot shows the iGeoportal interface with the three DeeGree WMS served layers mentioned before, combined into a single presentation. DeeGree iGeoportal has all standard functionalities of a webbased system.

IV. Conclusion

The SYNTHESYS NA-D 3.7 itinerary initiative has grown from a promising idea to a well-defined project. Within the time limits of the work plan, important goals were achieved :

* itineraries were situated in the broader context of biodiversity informatics

* An analysis of ABCD concepts usable for the itinerary project was made, and possible additions or changes to the schema, for better describing itineraries, were proposed.

* possible additional data sources have been identified and checked for validity.

* synergies with other data quality assessment tools have been explored.

* active collaborations with our SYNTHESYS partners have been built.

* the technical approach to the "itinerary" concept was thoroughly explored, several possible methods have been assessed.

* a formal (UML) description schema for "itineraries" and all related objects was made, clarifying internal and external communications.

* a testcase was found, which offers a broad range of data behaviour, while remaining limited enough to handle.

* tools and standards to be used were identified and tested

* a prototype algorithm for aggregating unit-level data was constructed, which yielded a lot of useful considerations and experiences.

* different possibilities for visualisation were tested.

Thus, a solid groundwork is laid for the further implementation of itineraries, interfacing with unit-level data and visualisation in the on line tools.

V. Further schedule for the itinerary project.

Further work on the itinerary project will follow the outline of the NA-D work plan :

June 1 2006 - September 31 2006

Implementing of a demonstrator website for visualising expeditions based on unit level data in the GBIF/BioCase cache. Deliverable : implementation of demonstrator.

* set-up of a server accessible from outside RMCA. Internal test servers are already running.

* installation of DeeGree WMS and iGeoserver software on the server.

* set-up of a limited dataset for testing the demonstrator.

* implementation of the algorithms in a website.

October 1 2006 - January 31 2007

Development of algorithms interfacing user unit level data with expeditions data (as in the cache and accessible through ABCD) for quality check. Deliverable : report on algorithms interfacing unit-level data with expedition data.

* technical analysis of expedition data / unit-level data interface

* development and testing of the interface algorithm prototypes

February 1 2007 - May 31 2007

Implementation in the interfaces of unit level data in the cache, with implementation of visualisation.Deliverables : implementation in interfaces, article for publication : GIS services.

VI. References

VI.1. References to literature :

Beller, A., Krumenacker, A. and Malicky, M. Report on available analysis tools and proposed choices for ENBI network. ENBI Report No: Wp10_D10.1a_06/2004. The Hebrew University of Jerusalem - Institute of Life Sciences, Evolution Systematics and Ecology, Israel. Biologiezentrum des OOE Landesmuseums, J.-W.-Klein Str. 73; A-4040 Linz, Austria.

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

Cox, S.J.D., Richard, S.M., 2005. A formal model for the geologic time scale and global stratotype section and point, compatible with geospatial information transfer standards. Geosphere, December 2005; v.1; no.3;p. 119-137.

Cox, S., Daisey, P.,Lake, R., Portele,C. and Whiteside, A., 2003. OpenGIS Geography Markup Language (GML) Implementation Specification v3.00, OGC document ref. no. : OGC 02-023r4, January 2003.

Cox,S.,Cuthbert,A.,Lake,R. and Martell, R., 2002. OpenGIS Geography Markup Language (GML) Implementation Specification v2.1.2. OGC Project Document No. : 02-069, September 2002.

Grønmo, R., Solheim, I. and Skogan, 2002. D. Experiences of UML-to-GML Encoding. SINTEF Telecom and Informatics, Forskningsveien 1, Pb 124 Blindern, N-0314 Oslo, Norway.

Guralnick, R. and Neufeld, D. Challenges building online GIS services to support global biodiversity mapping and analysis : lessons from the mountain and plains database and informatics project. Biodiversity Informatics,2,2005,pp.56-69.

Fitzke, J., Greve, K., Müller, M. u. A. Poth, 2003: Deegree - ein Open-Source-Projekt zum Aufbau von Geodateninfrastrukturen auf der Basis aktueller OGC- und ISO-Standards. Erscheint in: Geo-Informationssysteme 16 (2003), H.9, S. 10-16.

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

Andreas Krumenacker, Michael Malicky, Biologiezentrum des OOE Landesmuseums, J.-W.-Klein Str. 73; A-4040 Linz, Austria.

Martin R.C., 1998 "UML Tutorial", www.uml.orghttp://rd13doc.cern.ch/Atlas/DaqSoft/sde/uml/uml.html

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).

Meirte, D. 2005. Proposed schedule for the RMCA project, entitled

Providing itinerary related datasets and tools (for integration, visualisation and quality check). Internal RMCA/SYNTHESYS document.

Mergen,P.,Louette,M.,Snoeks,J.,de Meyer,M.,Meirte,D., 2005. The Royal Museum for Central Africa in the era of biodiversity informatics. Department of Zoology, Royal Museum for Central Africa. Leuvensesteenweg13, B-3080 Tervuren, Belgium. <poster>

Mergen, P., 2005(1). Mission Report, SYNTHESYS Network activity NAD 3.6-3.7 and GBIF.DE Coordination meeting at lat/lon GmbH, Bonn, Germany.

Mergen, P., 2005(2). Mission Report, SYNTHESYS Network activity NAD User Interface meeting Magyar Termeszettudomanyi Muzeum, 1088 Budapest, Ludovika tér 2.

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

Neale, S. and Pullan,M. SYNTHESYS report : User interface requirements analysis for the BioCase/SYNTHESYS biological information portal. Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh.

Portele,C.,2005 (1). Mapping UML to GML Application schemas. ShapeChange - Architecture and Description. Shapechange software documentation. Interactive instruments GmgH.

Portele,C.,2005 (2). Mapping UML to GML Application schemas. Guidelines and Encoding Rules. Interactive instruments GmgH.

Sinnott, R.W., 1984. Virtues of the Haversine; Sky and Telescope, 68(2): 159

Slik, J.W.F. & van Welzen, 2005. "Placing Biodiversity data on the Internet: Lessons learned" Original title : Report on responses to websites with selected feasibility studies regarding requirements for web-based access to digitised collection data. ENBI report No. WP13_D13.7_12/2005. Nationaal Herbarium Nederland, University of Leiden Branch, P.O. BOX 9514, 2300 RA Leiden, the Netherlands.

Slack, G., The First Comprehensive Survey of Northeastern Congo. Text on the AMNH website : http://diglib1.amnh.org/intro/intro.html.

Snyder, J.P., 1987.Map Projections - a Working Manual.US Geological Survey Professional Paper 1395, United States Gonvernment Printing Office, Washington D.C., pp.24.

Suzuki, J. and Yamamoto, Y., 1998. Making UML models exchangeable over the Internet with XML: UXF approach. Department of Computer Science, Faculty of Science and Technology, Keio University, Yokohama, 223-8522, Japan.

Torre, J.de la, 2005. SYNTHESYS NA-D 3.6. Report of existing GIS and software - Deliverable 3.6.1. Museo Nacional de Ciencies Naturales Consejo Superior de Investigaciones Científicas (CSIC)

Wieczorek, J., Guo, Q., Hijmans, R.J. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. Int.J.Geographical Information Science Vol. 18, No.8, December 2004, 745-767.

VI.2. references to the Web :

Access to Biological Collection Data (ABCD) :

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

American Museum of Natural History (AMNH) : http://diglib1.amnh.org/intro/intro.html

Apache Tomcat : http://tomcat.apache.org

Biological Collection Access Service for Europe (BioCase) : http://www.biocase.org/

BioGeomancer: http://www.biogeomancer.org/

CARPE project : http://carpe.umd.edu/index.asp

Catalogue of Life : http://www.sp2000.org/

Centro de Referência em Informação Ambiental (CRIA) : http://www.cria.org.br/

Committee on Data for Science and Technology (CODATA) : http://www.codata.org

DarwinCore : http://darwincore.calacademy.org/

DeeGree :

(Sourceforge) http://deegree.sourceforge.net/

(lat/lon GmbH) http://www.deegree.org/

Distributed Generic Information Retrieval (DIGIR) : http://digir.sourceforge.net/

DivaGIS : http://diva.riu.cip.cgiar.org/index.php

EarthSLOT : http://earthslot.org/iditarod/index.php/

European Network for Biodiversity Information (ENBI) :
http://www.enbi.info/forums/enbi/index.php

Global Biodiversity Information Facility (GBIF) : http://www.gbif.org/

GBIF Data Tester : http://gbif.sourceforge.net/datatester/

Geographic Markup Language (GML) : http://opengis.net/gml/

Google Maps : http://maps.google.com/

Google Earth : http://earth.google.com/

Iditarod : http://www.iditarod.com/

International Organisation for Standardization (ISO) : http://www.iso.org/

MySQL : http://www.mysql.com

MaNIS/HerpNet/ORNIS Georeferencing Guidelines :

http://manisnet.org/GeorefGuide.html

NASA Worldwind : http://worldwind.arc.nasa.gov/

National Botanic Garden of Belgium (NBG) :
http://www.br.fgov.be/PUBLIC/GENERAL/index.html

Open Geospatial Consortium, Inc (OGC) :

http://www.opengeospatial.org/

PgAdmin : http://www.pgadmin.org/

PostGIS: http://www.postgis.org

PostGreSQL: http://www.postgresql.org

QuantumGIS : http://qgis.org/

Structure of Descriptive Data (SDD) : http://www.diversitycampus.net/Projects/TDWG-SDD/Minutes/SDD-Introduction.html

SYNTHESYS : http://www.synthesys.info/

SYNTHESYS NA 3.6 : http://www.biocase.org//products/geo_services/index.shtml

SYNTHESYS Taxonomical Facilities (TAFs) :
http://www.synthesys.info/taf_contact_details.htm

Taxonomic Concept Schema : http://wiki.gbif.org/guidwiki/wikka.php?wakka=TCS

Taxonomical Database Working Group (TDWG):
ttp://www.nhm.ac.uk/hosted_sites/tdwg/

Unified Modelling Language (UML) :

UML models of GML Application Schemas:
https://www.seegrid.csiro.au/twiki/bin/view/Xmml/UmlGml

UMLEditor : http://www.umleditor.org

Web Map Service (WMS) : http://www.opengis.org/docs/01-068r2.pdf

Web Feature Service (WFS) :

http://www.opengis.org/docs/02-058.pdf

http://www.gbif.org/DataProviders/statistics1text

http://www.gbif.org/DataProviders/statistics1

http://www.gbif.org/DataProviders/statistics2

VII. Annex : RMCA workshop announcement

Announcement of the Workshop on the RMCA website (from : Mergen, 2006).

Meeting on GIS related projects around GBIF and TDWG

Workshop organized in the Framework of the EU project SYNTHESYS NETWORK ACTIVITY D (www.synthesys.info)

In the framework of its SYNTHESYS related task (Providing itinerary related services, http://www.biocase.org/products/geo_services/itineraries/), the Royal Museum for Central Africa is hosting a workshop on GIS related projects around GBIF (www.GBIF.org) and TDWG (www.TDWG.org).

The workshop will be attended by several international partners of the SYNTHESYS project and by GIS experts of GBIF-Germany and Spain. The status of ongoing tasks will be presented and discussed. A substantial part of the workshop will be devoted to evaluate several technical options and to brainstorming.

The workshop will take place the 22nd of February (10 to 17h) on the ground floor of the CAPA building (Leuvensesteenweg 17, Tervuren).

Please find in attachment the provisional program. If you or a representative of your Institution/department is interested in participating to this GIS workshop, contact us, until Fridays 17th.

Programme:

Topics:

- Georeferencing Collection Data GUI (Graphical User Interface)
- Map-like Visualization of georeferenced Collection Data
- GML Application Schema development for georeferenced Collection Data

Presentations:

Welcome and logistics (Michel Louette, Patricia Mergen*)*

1.) Status reports (along with brief demonstrations):
- Current Status (Javier de la Torre)
- Itinerary services + demo of existing visualizations (Bart Meganck, Danny Meirte, Patricia Mergen)
- GNOSIS Demo (Steven Smolders)
- Open the floor to GBIF Spain (Jesús Fernández Segovia, Ramón Pérez) ,

RMCA Geology (Pascale Lahogue) ,
- GBIF Germany (Jorg Holetschek), lat/lon (Jens Fitzke), University of Bonn (Christian Kiehle)

2.) Discussion

- Synergies between existing GIS related projects
- Compilation of the user requirements

3.) Future Work

- Remaining Tasks in the ongoing projects
- Plans for future projects