e-science and technology infrastructure for biodiversity data and observatories

Deliverable 5.3.1

Technical Construction Plan
European Strategies For Local Implementation -

"Thinking Globally, Acting Locally" -

Version 0.10

Cardiff University

Work Package 5a Construction Plan Strategy

Life Watch
e-Science and Technology infrastructures
for biodiversity data and observatories

EU Seventh Framework Programme (FP7) Infrastructures (INFRA-2007-2.2-01)

Contract Number 211372

Security (distribution level)	Private
Actual date of delivery	2009-12-31
Deliverable number	D5.3.1
Deliverable name	LifeWatch Technical Construction Plan European Strategies for Local Implementation
Type	Report
Status & version	Working Document v0.10 8th January 2010 for Rome Management Committee meeting
Number of pages	35
WP contributing to the deliverable	5a
Task responsible	Cardiff University
Other contributors	Olaf Bánki Wouter Los Axel Poigné
Author(s)	Alex Hardisty

Revision History

Version	Description	Changed	Date	Done by
V0.10	Edited to incorporate changes suggested by Poigné, Bánki, Los, etc.		2010-01-08	A.Hardisty
V0.9	Edited to match LifeWatch deliverable stanard		2010-01-07	A. Poigné
V0.8	Improvement of Title. New version for Stockholm MC meeting.		2009-10-09	A.Hardisty
V0.6	Split into separate documents and continued drafting, following review by WP5 and others		2009-09-07 onwards	A.Hardisty
V0.5	1st draft for discussion by WP5		2009-08-03	A. Hardisty
V0.4 and earlier	Private drafts			

Table of Contents

1 What is LifeWatch?

LifeWatch is an e-Science infrastructure to explore, describe and understand patterns in biodiversity and the processes that maintain biodiversity in space and time at the gene, species, and ecosystem level. The objective of LifeWatch is to establish a European research infrastructure to facilitate new frontiers in research concerning evolutionary, genetical, ecological, landscape and conservation aspects of biodiversity. Key elements of the infrastructure will consist of distributed observatories/sensors, interoperable databases, processing and analytical tools, computational capability, and computational capacity.

The innovative LifeWatch design offers integrated access to large-scale data resources, advanced algorithms and computational capability through a service-oriented architecture to support knowledge development on biodiversity from the genetic level to ecosystems and landscapes. Data mining, data analysis and modeling allows users to study patterns and mechanisms across different levels of biodiversity. The LifeWatch infrastructure provides scientific research teams with new collaborative environments by creating '*e-Laboratories*' or composing '*e-Services*'. They may share their data and analytical and modeling algorithms with others, while controlling access. LifeWatch enables "distributed large scale" and collaborative research on complex and multidisciplinary problems.

LifeWatch will be an open access infrastructure, offering a single portal for users from the scientific community, as well as policy makers and representatives of the private sector.

2 The Challenges Facing LifeWatch

There are several challenges facing the LifeWatch project as we specify and build functionality to meet the needs of potential users.

Firstly, there is a large gap between the ICT-based research practices in common/widespread use across the biodiversity research community today, and the vision of LifeWatch as a future interoperable 'e-Science' infrastructure offering collaboratory facilities to groups of scientists.

Secondly, the pace of ICT innovation is rapid, making it hard both to specify a stable platform that meets the needs of scientists and to "home in" on solutions with potential to achieve the LifeWatch vision. Almost every day sees new announcements and new leaps forward in what is possible. Researchers across Europe are quick to experiment with such new technologies, creating whole new tools and data platforms that depend upon them. Often this is done without thought for future-proofing and interoperability with other components in a wider landscape. The result is a jumble of heterogeneous tools and data resources, with each 'owner' championing their own approach as 'the best one'. Faced with such an existing landscape of heterogeneous data resources and tools, together with a long-term vision for what is needed or desirable, this presents a significant integration problem as LifeWatch prepares to construct *infrastructure*[1].

Progress towards solving the first challenge requires engagement with open-minded scientists willing to engage, explore and progress. Scientists who can evangelise; who can put into practice a new modus operandi for biodiversity research; not only expert in their own domain but also enthusiasts for and technically competent with the new technologies.

Progress towards solving the second challenge is in part accomplished by the LifeWatch Reference Model (LifeWatch-RM). LifeWatch-RM is a mechanism for making technology independent design decisions now, for the basis of the preparatory work, which can be instantiated later-on using specific technology approaches.

These approaches can be supported both by an appropriate technical governance model, based on the meritocratic approach of the Apache Software Foundation, and the strategy set out in the present document. We call this "European strategies for local implementation" or "Thinking globally, Acting locally". By this we mean that the role of LifeWatch is to devise and apply European level strategies to bring about

[1] "Infrastructure" is the permanent elements needed to create an internet/web-based inter-organisational system that links personal and institutional systems (data resources and analytical tools) for biodiversity research.

community collaboration and interoperability, which can be supported or complied with through implementation at the local level. In other words: We *think* at the European level to bring about technical cohesion and interoperability, and to set direction, determine priorities, manage performance, and hold to account. We *engage* the community and act at the *local* level to implement technical elements.

This approach explicitly recognises the importance of the many small-scale communities of collaboration that exist in biodiversity research at both national and European levels[2], whilst at the same time recognising the need for over-arching guidelines to achieve interoperability and greater collaboration.

In what follows we first create the "big picture" strategy (European Strategies, Thinking ‘globally’), followed by the related elements that have to be enacted at the local level (Local Implementation, Acting ‘locally’).

2 And, actually also at the global scale too.

3 LifeWatch 'Unique Selling Point' and 'Value-Added' Benefits

The 'Unique Selling Point' (USP) of LifeWatch is that it is an ***infrastructure***. That is to say: It is the permanent elements needed to create an internet/web-based inter-organisational system that links personal and institutional systems (data resources and analytical tools) for biodiversity research, supported by appropriate human resources providing assistance to users.

In this sense, the aspirations for LifeWatch are to:

- Provide a full range of functions across multiple scales. These include functions for: data gathering and generation; data management, integration and modelling; and to support diverse applications. These functions will integrate across genomic; organism; species; ecosystem; and landscape scales;
- Support globally unique identification of biodiversity resources, both physical assets and biodiversity concepts;
- Support workflows as the paradigm for accomplishing specific research tasks involving the transformation, processing and analysis of data;
- Support mechanisms of provenance to permit tracing of data and workflows for reproducibility of scientific analysis, and tracing of data re-use and citation;
- Be compliant with the requirements of the INSPIRE Directive;
- Semantically interoperate heterogeneous data resources.

The 'value-added' benefits of LifeWatch are that it:

1. Enables discovery and access to a wide variety of data - genetic, ecological and abiotic - to support biodiversity research and policy;
2. Supplies the infrastructure for managing, merging and manipulating data from multiple sources;
3. Provides taxonomic support e.g., authoritative species lists and taxonomic classifications, digitisation-on-demand;
4. Provides support for the spatial mapping of data; and,
5. Supports sharing of workflows, collaboration and community-building.

In terms of strengthening the LifeWatch USP and value-add, we draw on key insights from the e-Biosphere conference, held in London 1-3 June 2009.

Firstly, "*it is not sufficient just to organise biodiversity data. The key is to link it to other socio-economic data for policy-making purposes.*"

Secondly, we MUST "*link biodiversity informatics to ecosystem services"* that are relevant to a broad range of governmental policy questions and commercial interests[1]. LifeWatch must deliver benefits to these users through scientists being able to tackle the societal questions these users ask.

Thirdly, "*linking information across scales and disciplines*" and also across heterogeneous sources, sites and organisations could be seen as the USP - provided we are able to demonstrate it. We must remember that a USP does not come into existence until we can actually demonstrate it and establish some credibility. Therefore, we have to act locally to bring it about, primarily through the interoperability framework represented by the LifeWatch Reference Model.

In LifeWatch data, services, tools and facilities will be distributed among the participating countries. To become fully operational will take 5 years. Using a phased approach LifeWatch will release constant small increments in functionality. Driven by need, LifeWatch will take its requirements input from the User Platform and the Data Resources Platform. The LifeWatch roadmap[2] plans the first 3 years of construction. Using the European strategies for local implementation approach, we expect to see "islands of infrastructure" emerge and fuse across Europe to deliver a pan-European interoperable infrastructure to meet the needs of the community. The local implementation strategies are the key mechanisms to make this happen. The European strategies (i.e., thinking globally) are the mechanisms to ensure that local action occurs in a coordinated and guided way.

---

[1] It is these sectors that have the money to pay for LifeWatch science.

[2] Currently (January 2010) in preparation.

Blank Page

4 European Strategies, Thinking ‘globally’

"European Strategies" are concerned with two aspects:

1. Building and governing a pan-European interoperable infrastructure that crosses national and institutional boundaries; and,
2. Seeking to encourage, from a top-down perspective, alignment of multiple separate and autonomous initiatives (individual, research-group, institutional, etc.) to an emerging European (and global) framework of biodiversity infomatics rules designed to bring about interoperability in a heterogeneous community.

The following sections consider some of the aspects having a European rather than a local perspective.

4.1 Planning objectives and overall construction strategy

LifeWatch prepares to construct and operate an infrastructure that:

- Is “fit-for-purpose”
- Integrates “external resources” provided by institutions & networks concerned with biodiversity research
- Offers “Capabilities” to ensure early engagement and attract widest community of users
- Is easy to use at different levels of knowledge and understanding in both science and policy domains
- Covers not only the scientific usage cases, but also the operational and technical constraints
- Is “Open” and based on industry standards
- Uses existing technological solutions
- Takes heterogeneity in its stride
- Is operational at the earliest opportunity
- Is staged - not everything available on ‘day 1’.

Our strategy for construction revolves around the concept of “loci” or “centres” offering “services” to users. Loci can be:

- ICT focussed i.e., computerised loci;
- Human focussed i.e., service centres; or,
- A combination of the above.

NOTE: The concept of a locus or centre in LifeWatch is one of a focal point or node offering capability to serve the LifeWatch community.

Some loci / centres may be data resources owned and operated by Data Resource Owners. Some loci / centres may be tool resources owned and operated by Tool Owners. Some loci / centres may be ‘Service Centres’ offering human oriented services, such as specialist advice and support.

All of these kinds of loci / centres may offer specific services to LifeWatch users. Loci / centres can be distributed across participating countries and implementation is a collective partnership between LifeWatch, the national initiatives and individual institutions around shared strategies. These include: strategies for sharing and pooling resources; strategies to capitalise on inhouse and community capabilities; and strategies for European and national leadership. The basis on which services are offered by Loci can be negotiated with the LifeWatch ERIC (see 4.2 below). LifeWatch will offer several model Service Level Agreements to suit a range of different circumstances.

LifeWatch will provide:

- Organization
- Technical direction & governance;
- Core infrastructure;
- Management of the “Product”; and,
- Community support.

4.2 The LifeWatch organization

LifeWatch will be a distributed research infrastructure, where the entities of the LifeWatch organization and the services that the LifeWatch organization provides will be spread over Europe. Several entities important for LifeWatch could also be outside the LifeWatch legal framework, and in these cases formal linkages are required. Countries will be involved in LifeWatch through legally binding to the ERIC Governing Board (see Figure 1). A minimum of three countries is required in the ERIC to proceed with the construction phase. In the construction phase the LifeWatch organization will be established. The LifeWatch organization will be a managing organization consisting amongst others of a front office, an executive management office, and several service centres for technical support, research and strategy development, etc. This LifeWatch organization will engage through formal agreements and operational relations with independent national centres (this can be existing organizations, government agencies, ect.). At the same time there could be relations between the countries and the independent national centres through the ERIC. So the relation of the LifeWatch organization with national independent centres is plural. The LifeWatch organization will also formally engage with the user and data/resource provider community. Individuals may either engage with LifeWatch through the community or through the independent national centres (see Figure 1)[1].

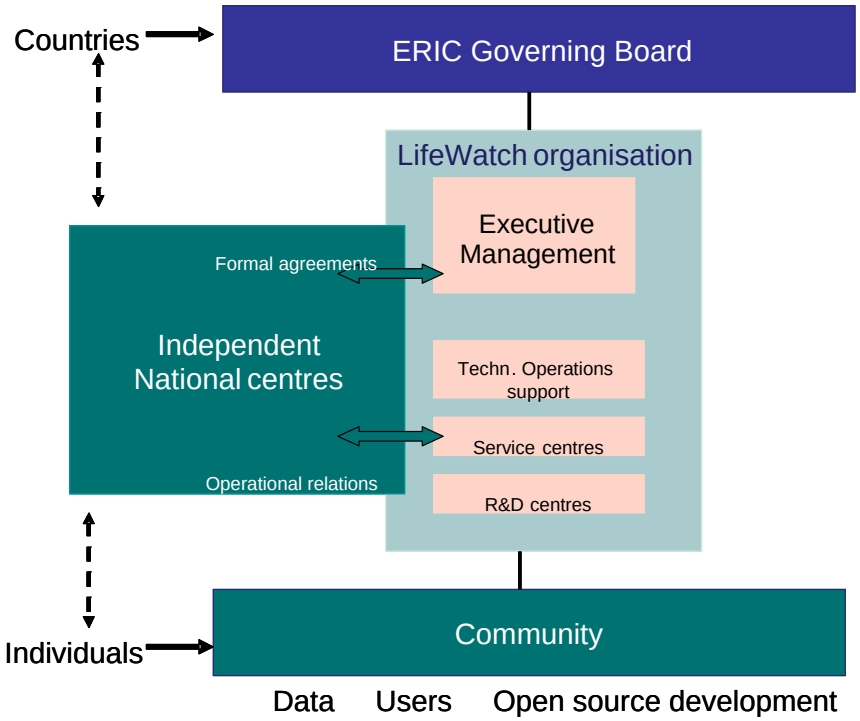

Figure 1. The LifeWatch organization and its relations

4.3 The Architecture of the LifeWatch ICT infrastructure

The LifeWatch Architecture, as described by the LifeWatch Reference Model (LifeWatch-RM) is the technical basis of European strategies for local implementation. It is a set of guidelines for encouraging the development of the infrastructure in a certain direction, consistent with the vision. LifeWatch-RM is a mechanism for making technology independent design decisions now, for the basis of the preparatory work, which can be instantiated later-on using specific technology approaches.

[1] For additional information on the LifeWatch organization see the LifeWatch masterplan

In summary, Figure 2 below illustrates the main elements of the LifeWatch Architecture.

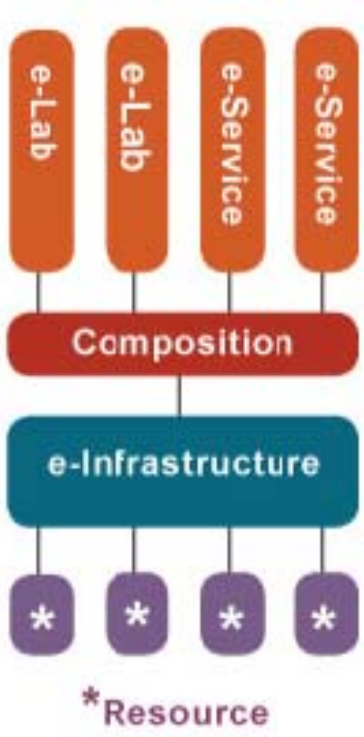

Figure 2: Main elements of the LifeWatch Architecture

Four layers are depicted in Figure 2, as follows:

- The *Resource Layer* contains the specific resources, such as data repositories, computational capacity, sensor networks / devices and modelling / analysis tools, that contribute to the LifeWatch infrastructure. The primary components at this layer are the biological (e.g. species records) and abiotic data from sites. Additional components include catalogue services (e.g., taxonomic checklists or gazetteers), analysis tools, and processing resources.

- The *e-Infrastructure layer* provides mechanisms for sharing the specific resources as generic services in a distributed environment, spread across multiple administrative domains. This includes systems for identifying, accessing, and processing resources located within multiple administrative domains, uniform security and access protocols, and ontologies for semantic integration.

- The *Composition Layer* supports the intelligent selection and combination of services in order to complete tasks. This includes workflows, semantic metadata for the discovery of components and the storage of additional attributes such as provenance and version information. Viewed from the perspective of the biodiversity scientist, the Composition Layer consists of a wide range of *application services* or *biodiversity research capabilities* that can be selected and arranged in an infinite number of ways to conclude specific analytical and experimental tasks.

- The *User Layer* provides domain-specific presentation environments for the control and monitoring of tasks and tools to support community collaborations. This includes a LifeWatch portal incorporating discovery and workflow management tools and offering a single point of access for all users.

The reader is referred to the LifeWatch Reference Model document for more detailed information (see on www.ve-forum.org).

4.4 MasterPlan and Jira Database

LifeWatch uses a database called 'JIRA' (http://jira.lifewatch.eu/) to keep track of the multiple projects, components and issues associated with construction of the infrastructure. Together with the 'MasterPlan' the content of this database describes, with costs and risk assessment, everything to be constructed to create the LifeWatch infrastructure.

4.5 Showcases, services and pre-defined workflows

4.5.1 Showcases

Showcases are synonymous with community applications ("e-Labs / e-Services" in the terminology of the architecture). Showcases represent particular thematic sub-domains of biodiversity research and policy-making. We use showcases for two purposes:

Firstly, showcases are examples to explain to potential users of LifeWatch, the kinds of applications and tasks that users should be able to accomplish with the infrastructure when it is built. Of course, the most interesting showcases for us are the high-value ones addressing biodiversity problems of major economic and societal significance that it is not possible to achieve today. These can demonstrate to Member States the value of investing money to support the construction and operation of LifeWatch.

Secondly, we use the showcases to derive the underlying functions (such as calculating the distribution range of a species) that are common to multiple applications. By building LifeWatch to offer these functions we provide the means to implement particular known showcases. By providing workflow mechanisms that allow the functions to be "arranged and executed" in a multiplicity of ways, we also have the ability to support use-cases that are unknown to us at the present time. Over time, we will gradually extend the list of supported functions so that a wider and wider variety of tasks can be accomplished.

In the LifeWatch Reference Model we refer to these functions as "biodiversity research capabilities". We realise them in the ICT infrastructure as "services".

It makes sense to consider services in thematic groups to support particular application sub-communities. This eases both "product management" and co-ordination of pilot / national / local projects as we can align sub-community groups of interest to particular thematic groups of services. Of course, there are some fundamental services that will be common to many sub-communities and we group into these into several categories of more general services.

The current list of showcases is as follows:

1. Biodiversity Richness Analysis and Conservation Evaluation
2. Biological Valuation Map
3. Automated Retrieval
4. Past Behaviour and Future Scenarios
5. Bioclimatic Modelling and Global Climate Change
6. Phylogenetic Analysis and Biogeography
7. Ecological Niche Modelling
8. Urban Development and Biodiversity Loss
9. Renewable Energy Planning
10. Hierarchical scaling of biodiversity in lagoon ecosystems along environmental gradients
11. Bird Strike Monitoring
12. Earth Observations

These are elaborated further in a separate document ("Showcases", Deliverable D5.1.2).

To date, we have 12 showcases that are more-or-less based around what we conceive as possible at the present time. This is not sufficient and we need some more far-sighted ones like the following examples[2]:

1. agri.be: agricultural management application balancing production of quality farm produce with optimisation of farm biodiversity

[2] Source: Wouter Los, UvA.

2. Identify small animal / plant populations with stagnating genetic variability. Determine if they can be restored by linking to other populations nearby (i.e., where to create ecological corridors)?
3. What is the likelihood of a particular horizontal symbiont / parasite transfer (epidemic) among a set of closely related species?
4. Habitat classification for species and classification of distinct ecosystem types are important tools for landscape management. What is the impact over time of external pressures (e.g., urban sprawl) on any classification?
5. Vegetation cover and plant root structures are important factors for hydrology. What scenarios, stemming from climate change, result in the collapse of a particular hydrological system and what mitigating strategies may be effective? Note that this kind of investigation could be part of a showcase on larger river catchments e.g., the Donau catchment or the Rhine catchment.

According to need, some show cases are likely to be implemented in the LifeWatch infrastructure as pre-defined workflows (also known as “biodiversity themed research tasks”; also known as “applications”) (see 4.5.4).

4.5.2 General Services

Even with the limited analysis so far, we can already identify more than 60 common biodiversity research capabilities ((for example, the ability to plot the distribution of something on a map) that are necessary to achieve the 15 showcases listed above. These include multiple different capabilities for, for example:

- Data discovery
- Data access
- Data processing
- Geospatial processing
- Modelling
- Visualization
- Support for users and user groups.
- etc.

Generally speaking, these capabilities fall into a category of services that we call “LifeWatch Basic Services”.

LifeWatch Basic Services are the common services necessary for most applications of LifeWatch. They are made up of services from a number of sources:

- the Biodiversity Research Capabilities;
- Open Geospatial Consortium (OGC) services relevant for LifeWatch (e.g., to support compliance with the requirements of the INSPIRE Directive);
- the fundamental ORCHESTRA Architecture Services.

4.5.3 Thematic Services

On top of these capabilities, we can then identify (potentially, hundreds) of capabilities to support particular sub-domains of biodiversity work. At present, we foresee several thematic service categories, as follows:

- Biodiversity Valuations
- Citizen Science and Observations
- Observatory Tasks
- Phylogenetics
- Species Dynamics and Distribution
- Species Richness and Ecosystem Services.

4.5.4 Pre-defined workflows

Services, as described above, in conjunction with workflow tools offer users the ability to create and execute their own workflows in order to analyse and model the data of their interest in relation to a specific research question. However, not all users will require such sophisticated flexibility. Many users will be

content to make use of pre-defined workflows. These are a suite of tailor-made workflows for frequently requested analytical and modelling methods.

There is an outstanding issue concerning what pre-defined workflows should be supported by LifeWatch in its early operational phase.

4.6 LifeWatch Interoperability

Technical interoperability in LifeWatch is partly achieved by the technical framework and architecture (see 4.10). Other details will be specified in an Interoperability Plan, to be derived from LifeWatch Reference Model elements and the needs of the national projects. It is likely to contain parts dealing with interoperability at the following levels:

- Service Model
- Information Model
- Biodiversity Information Standards (TDWG) - See document: www.tdwg.org.

NOTE: The Atlas of Living Australia (ALA) contains some useful information about the support of TDWG standards here: http://www.ala.org.au/tools/data-standards.html.

4.7 Relations to other organisations and initiatives

We need to be clear about the nature of the relationship between LifeWatch and each of the other relevant organisations and initiatives. Is the organisation / initiative a partner of LifeWatch? Is it a consumer of LifeWatch services? Is it a supplier of something to LifeWatch?

Each relationship is different and will have to be individually negotiated to arrive at a bi-lateral agreement. Such negotiation will have legal, financial and technical components. This is mainly something that will need to be done during the early part of the Construction Phase, but which may start in the rest of the Preparatory Phase (e.g., as for EEA).

Organisations and initiatives with which LifeWatch has to establish relations include:

- Aurora Borealis - the new European Research Icebreaker ship for use in arctic research;
- Species2000/ITIS Catalogue of Life
- European Environment Agency (EEA)
- GBIF

NOTE: GBIF is moving into its operational phase. Currently (June 2009) GBIF provides access to 8080 datasets from 294 Data Resource Owners. It seems clear that a lot of what GBIF is doing is relevant to LifeWatch, but how? It seems to be relevant at the Resources level and the lower parts of the Infrastructure level of the LifeWatch architecture. Alignment to how GBIF approaches matters at these two levels could result in considerable savings for LifeWatch, allowing us to focus effort on the additional value offered by the higher levels of our architecture. Things that could be relevant at the early stages of LifeWatch construction include:

- IPT: Integrated Publishing Toolkit. As a means of bringing External Source Systems on-line. The GBIF IPT as presently designed is probably not of sufficiently general purpose to meet the needs of all LifeWatch Data Resource Owner and data categories. LifeWatch should: a) consider the appropriateness of entering into an arrangement with GBIF to deliver a more general purpose tool that also meets LifeWatch needs b) determine whether IPT is being developed as an open-source community tool, which LifeWatch developers could enhance to meet LifeWatch needs.
- GBRDS: Global Biodiversity Resource Discovery Service. Effectively, GBIF's UDDI registry for biodiversity resources. To what extent can it accommodate all categories of data resource of interest to LifeWatch.
- GNA: Global Names Architecture ...GUIDs, GNI, GNUB - ultimately all data sources will be represented in the GNA.
- GEO-BON
- GMES

- iBOL
- ICOS
- NOHA
- SEIS

4.8 Governance

A technical governance regime needs to be worked out that will help to integrate the 'European Strategies' and 'Local implementations' aspects together.

One possibility we should consider is the OSS Watch meritocratic governance model, which itself is a derivation of the model used by the Apache Software Foundation.

In this model, individuals assume "roles" that have different levels of ability to affect project outputs and there is both a consensual decision-making process, and where necessary a voting process, for governing how the project develops.

Using such an approach allows increments in functionality to be released early and frequently. It involves users in the development process and allocates control to those individuals with sufficient knowledge to be able to exercise it appropriately.

Coupling this approach with the idea to find a small group of biodiversity scientists really willing to engage with LifeWatch would start to address the main challenges identified.

4.9 Technical Construction of the infrastructure

The infrastructure is the essential ‘central’ elements that LifeWatch needs in order to be able to function across Europe. Typically, the elements of the infrastructure are devolved and distributed. They may be outsourced to specific institutions to operate on behalf of LifeWatch. It is possible that some such elements will exist in a hierarchy but this needs to be better determined. The sections below describe the core components we currently envisage as necessary.

4.9.1 Coordination & management; central staff

Development and deployment should be overseen by a core LifeWatch team who are employees of the LifeWatch legal entity. This team should consist of:

- A Technical Programme Director
- 4 x Systems Architects
- 2 x Senior Software Engineers
- 4 x Software Engineers
- 2 x Infrastructure / Technical Support Engineers
- 1 x Project Manager
- Administrative support, incl. PA to the Technical Director.

Note, however, that it is not necessarily the case that this team of staff will undertake all aspects of software development. See also 0 below.

Technical operations support will be provided by the central staff to Data Resource Owners and Tool Owners. Central staff will also provide 2nd line support to users in cases where 1st/front-line support provided through the Service Centre is insufficient to solve a users’ problem. Central staff will broker 3rd line support from appropriate Data Resource Owners or Tool Owners.

The Programme Management Office will provide project management capability to oversee construction of the technical infrastructure.

4.10 Technical framework and architecture

Is provided by i) the LifeWatch Reference Model (ref to doc) ii) LifeWatch Service Networks and iii) the processes to support compliance.

Compliance processes will be needed for Data Resource Owners (Data Resource Owner “admission” process), Tool Owners (Tool integration process) and for software development (development process).

4.10.1 The LifeWatch Reference Model

The LifeWatch Reference Model (LifeWatch-RM) is the first step in providing the technical guidance necessary to achieve pan-European interoperable infrastructure.

In essence, the approach it takes is to create a separation between the more concrete specification of technology and the abstract specification of technical capabilities that are necessary to meet the LifeWatch goals. Such an approach gives 3 freedoms over the long-term:

1. Independence from constraints arising from evolution of technology;
2. The possibility of extending technical capabilities to support new requirements in the future:
    a. Functionalities expressed as services
    b. LifeWatch ‘nodes’ / ‘centres’ are service providers
3. Support for thematic extensions.

4.10.2 LifeWatch Service Networks

A LifeWatch Service Network (LSN) is a set of networked hardware components and service instances that interact to serve the objectives of an application. Service instances are the basic units of functionality.

An LSN corresponds to the concept of an ORCHESTRA Service Network (OSN) in the ORCHESTRA framework. An ORCHESTRA tutorial[3] gives a simple example of how to build an OSN.

4.10.3 Data Resource admission process

LifeWatch aims to present a consistent image of data resources to its users. The capabilities and the quality of the data provided by the LifeWatch Data Resource Owners (DRO) is a key to presenting that image.

The concept of a "LifeWatch certified Data Resource" is an important branding element for marketing purposes. To provide this consistent image, LifeWatch must encourage DROs to meet particular requirements and standards, both technical and non-technical. But LifeWatch also relies on the presence of many DROs. It is of no benefit to LifeWatch and LifeWatch users if DROs find the LifeWatch requirements too onerous, with the consequence that they fail to join the LifeWatch infrastructure.

To ensure successful integration of DROs into the LifeWatch infrastructure, it is necessary to ensure two aspects: Firstly, the practical needs of and constraints on DROs must be well-understood by LifeWatch. Secondly, taking these needs and constraints into account, LifeWatch must be able to provide a clear and unambiguous statement of the requirements that a DRO must meet in order to be "admitted" to the LifeWatch infrastructure.

To become associated with LifeWatch, DROs will need to commit to a "Service Level Description" (SLD), detailing their provision of data resources and any constraints on the use of those resources. While it is important not to discourage eager participants from joining, LifeWatch must ensure that participants can contribute meaningfully to an integrated system. Hence, the SLD template must be both rigorous and flexible in order to successfully handle the DROs' differing expectations and capabilities.

---

3 http://www.eu-orchestra.org/TUs/OSN/en/html/index.html

In addition, DROs will have to commit to proceeding through a process of integration with the LifeWatch infrastructure, or “data resource admission” as we propose to call it. This admission process will consist of a sequence of steps that begins with identifying the data to be contributed, defining and implementing the technical interfaces and SLD to make that possible, and culminating in a period of rigourous testing to ensure that the DRO can meet the commitments set out in the SLD.

Further specification of this admission process is presently (January 2010) the subject of a specific sub-project within the Preparatory Phase that will result in:

- A selection of policy considerations to determine the admission of DROs to the LifeWatch system.
- A Service Level Description (SLD) template to define candidate "service levels" that can be expected from DROs. The SLD template can be tailored to the particular requirements of an individual DRO as part of the admission procedures. The template will provide potential DROs with a detailed description of the benefits and commitments required for them to participate in the LifeWatch infrastructure. A DRO can be subsequently monitored against the requirements of the agreed SLD to ensure that the high standards for LifeWatch data quality are maintained throughout the operational phase.
- A series of procedures for "admitting" (i.e., signing up, integrating and testing) a DRO to the LifeWatch infrastructure.

4.10.4 Tool integration process

LifeWatch aims to present a consistent image of available tools to its users. The capabilities and the quality of the tools provided by the LifeWatch Tool Owners (TO) is a key to presenting that image.

The concept of a "LifeWatch certified Tool" is an important branding element for marketing purposes. To provide this consistent image, LifeWatch must encourage TOs to meet particular requirements and standards, both technical and non-technical. But LifeWatch also relies on the presence of many TOs. It is of no benefit to LifeWatch and LifeWatch users if TOs find the LifeWatch requirements too onerous, with the consequence that they fail to join the LifeWatch infrastructure.

To ensure successful integration of TOs into the LifeWatch infrastructure, it is necessary to ensure two aspects: Firstly, the practical needs of and constraints on TOs must be well-understood by LifeWatch. Secondly, taking these needs and constraints into account, LifeWatch must be able to provide a clear and unambiguous statement of the requirements that a TO must meet in order to be "admitted" to the LifeWatch infrastructure.

To become associated with LifeWatch, TOs will need to commit to a "Service Level Description" (SLD), detailing their provision of tools and any constraints on the use of those. While it is important not to discourage eager participants from joining, LifeWatch must ensure that participants can contribute meaningfully to an integrated system. Hence, the SLD template must be both rigorous and flexible in order to successfully handle the TOs' differing expectations and capabilities.

In addition, TOs will have to commit to proceeding through a process of integration with the LifeWatch infrastructure. This tool integration process will consist of a sequence of steps that begins with identifying the tool to be contributed, defining and implementing the technical interfaces and SLD to make that possible, and culminating in a period of rigourous testing to ensure that the TO can meet the commitments set out in the SLD.

Further specification of this integration process is still required and will result in:

- A selection of policy considerations to determine the integration of tools to the LifeWatch system.
- A Service Level Description (SLD) template to define candidate "service levels" that can be expected from TOs. The SLD template can be tailored to the particular requirements of an individual TO as part of the integraton process. The template will provide potential TOs with a detailed description of the benefits and commitments required for them to participate in the LifeWatch infrastructure. A TO can be subsequently monitored against the requirements of the agreed SLD to

- A series of procedures for integrating (i.e., signing up, integrating and testing) a TO to the LifeWatch infrastructure.

4.10.5 Development process

The long-term development model will be an open community development model, such as that used by the Apache Foundation.

However, in the first years of the construction phase there will be a significant amount of development and deployment work to be undertaken. This will be best accomplished on the basis of a commercial partnership contract placed by LifeWatch with an appropriate computing software and services provider. The model here will be for the commercial provider, operating within the context of the foreseen community model, to undertake detailed development and deployment work under the supervision of and in partnership with the core LifeWatch team. The model will need to ensure that appropriate knowledge transfer to the LifeWatch team occurs throughout the process. One possibility is to imagine an overarching "framework contract" with individual “develop and deliver” projects within it.

We need to retain flexibility to source specific components from other vendors as necessary and to pass them into the framework contract for integration.

As a general principle, open source software will be used. However, where appropriate software does not exist, proprietary software may be chosen on a case- by-case justified basis, at the decision of the Product Management Board (see 4.17).

4.11 Core ICT Infrastructure (Hardware/Software)

4.11.1 Datasets catalogue (registry)

This is the catalogue containing details of Data Resource Owners (publishers and individual providers) and datasets known to the LifeWatch infrastructure with a pointer to a URI for where to retrieve them. The catalogue will be sub-divided according to the established sub-domains of data: terrestrial ecology, marine ecology, taxonomy, molecular biology, genomics, and abiotic data. Other categories can be added as needed.

4.11.2 Services and tools catalogue (registry)

This is the catalogue containing details of (processing) services and tools known to the LifeWatch infrastructure with a pointer to a URI for where to access them. The catalogue will be sub-divided on a "yellow pages" basis according to the types of services and tools indexed.

4.11.3 Provenance and citation tracking repository

The Provenance and Citation Tracking Repository stores provenance meta-data to allow:

- Tracing the origin of data and workflows used/cited within LifeWatch (backward direction towards point of origin); and,
- Tracking the use and re-use of data and workflows provided into the LifeWatch infrastructure (forward direction towards point of academic publication).

The first mechanism is needed in order to reliably reproduce results claimed by publication. The second mechanism is needed in order to keep track of who is using a particular provider's intellectual property.

4.11.4 Annotation repository

The Annotations Repository stores annotation metadata. Typically, annotation metadata consists either of comments and notes about a resource (dataset, service, tool, etc.) added by users of that resource or of one or more edited dataset records. Comments and notes serve as quality statements about the resource

or highlight problems and shortcomings about the resource. Edited records are proposals to change the data source. When a resource is "discovered" the user is presented with related annotations from the Annotation Repository.

4.11.5 Portal

The LifeWatch Portal is a 'one-stop shop' for accessing the LifeWatch Infrastructure. It is the SINGLE POINT OF ACCESS to the LifeWatch infrastructure.

In the Portal a user can find all available resources such as data, workflows, tools, etc., and make use them through 'LifeWatch Portal Utilities'. Portal Utilities are self-contained mini-web-applications providing a specific functionality, and which are not intended to be used stand-alone but to be embedded into web applications. Common examples of portal utilities are widgets, gadgets and portlets. Portal Utilities provide user interfaces to basic capabilities of the LifeWatch infrastructure and can be combined into a complex Application to solve particular goals. The Portal provides a common platform for the exchange of messages and events between the utilities within an application.The Portal will serve as single entry point to all LifeWatch applications, integrating different utilities, for instance, exposing each application as a portlet or tab within the portal web site.

Vera's "Mock-up" document makes some suggestions for what this might look like. Further work-breakdown has been done in JIRA based on that.

Note: Likely to be based on the idea of portal server(s) separated from application server(s), in conjunction with social collaboration / networking tools, per (for example) the IBM Websphere model.

4.11.6 (Access to) Computational resources

Access to computational resources is required to complete computationally intensive tasks such as analysis, modelling, and simulation. LifeWatch will rely on European computational capacity provided by the European Grid Infrastructure (EGI) and the Partnership for Advanced Computing in Europe (PRACE).

4.11.7 Security (Authentication, Authorization, Accounting)

LifeWatch infrastructure is a federated information system in which users should be able to explore the federated systems/resources with as few barriers as possible. Once the user's credentials have been are established, the user should be able to (within the constraints of appropriate authorization) seamlessly access resources provided by different, providers, communities, etc. The requirement therefore, is for single sign-on with portability / federation of identity information across otherwise autonomous security domains. The security infrastructure must support both individual users without affiiliation to an institution (i.e., an OpenID type of approach) and institutionally affiliated users (i.e., a Shibboleth type of approach).

4.11.8 Semantic mediation framework

Semantic mediation will be a cornerstone for interoperability between resources of the LifeWatch infrastructure. The need for semantic mediation arises in several areas:

- Data and Service Discovery: Discover data and services based on, for example, specific domain ontologies;
- Data Mediation: Processing data based on its semantics even if the data is provided by different data models;
- Data Fusion: Combining data from different sources;
- Data Interpretation: Multiple data models and heterogeneity at the data level itself, perhaps arising from differences of professional opinion, makes interpretation more challenging;
- Service Integration: Chaining of services often needs transformation of data when passing data from one service to another; and,
- Workflow identification: Discovery of workflows that, e.g., may help to solve a particular modelling problem.

The approach proposed at present is to base the Semantic Mediation Framework on the UMLS approach of metathesaurus and semantic network. Like the medical field, we already have a number of distinct emerging vocabularies in the biodiversity domain and we also will want to link in data sources from outside of that direct domain (e.g., socio-economic, genomic, etc.). The UMLS approach will be supplemented with the extensibility mechanisms of SERONTO (ref) and the observation ontology aspects from OBOE. This construction issue will involve creating and deploying a LifeWatch MetaThesaurus and a LifeWatch Semantic Network.

4.11.9 Virtual collaborative environments

January 2010 - Urgent further work is needed to adequately describe and scope this component, based on the work of the BTCN.

Construction of digital tools to create personalized workspaces

Creation, management and dissolution of temporary collaborative networks / virtual organisations.

See Hamideh's report for details. Translate into construction issues.

4.12 Data resources

4.12.1 Wrapper development kit

Required for wrapping and admission of data resources.

4.13 Increased and targetted data generation

Central coordination and interoperability development

Requires: Instrumentation, common format development, site-level management, coordination costs, decentral administration, housing and overheads

- Marine sites
- Sensor data resources and networks
- Systematics collections - systematics data collections and reference lists
- Taxonomic backbone - structure and set of (European) standards for the taxonomic classification of all European databases of biodiversity data. Both directed at the long-term maintenance of crucial EU-nomen services and the integration of Global Species databases into the Cybertaxonomy platform. Issues are a.o. management of IPR taxonomic data, Global species databases (GSDs) and Global Names Architecture contribution
- Terrestrial and freshwater sites

4.14 Data processing (incl. Analytical & Modelling tools)

4.14.1 Processing capabilities ("Toolkit")

(Interoperable) environment for statistical and other software

4.14.2 Workflow generation and management

- Workflow composition tools
- Sharing of workflows
- Workflow library/registry, potentially with semantic interoperability determination

4.14.3 Integration of existing tools

- Service wrapping kit
- Action plan for biodiversity informatics tools integration
- Procedures for integration of new tools

4.14.4 Reliability, Availability and Servicability

Software systems and back-up hardware systems, redundancy, performance

4.15 Intellectual Property Management

There is a need to work out, legally, the different kinds of licensing agreement that will be needed between LifeWatch and its Data Resource Owners / Tool Owners and between users and LifeWatch.

Four possibilities will be needed, since the licensing agreements could be on data / tools owned by the providers or on data / tools owned by LifeWatch itself (if it collects or ends up owning data/tools)

- Directly between providers and users;
- Between providers and users having LifeWatch as intermediary
- Between providers and Life Watch
- Between LifeWatch and users

Once the needed licensing agreements have been worked out, the electronic means of entering into and enforcing these agreements must be put in place.

4.16 Innovation Lab (Research and Development)

We expect that there will not necessarily be solutions available to all technical problems associated with the LifeWatch aspiration when construction commences.

Therefore, during the construction phase we need to make provision to undertake some ongoing research activities.

Items that appear on the list of proposed activities at present are:

1. Semantic mediation framework
2. Sensor networks for:
    a. Climate change induced biodiversity change
    b. Organism distribution assessments

3. LifeWatch Grand Challenge: To be able to undertake ecosystem impact assessment and prediction at the landscape scale in response to requirements coming from:

- Large-scale civil engineering proposals e.g., Venice Lagoon, Severn Barrage; and,
- The foreseen effects of climate change.

4.17 Product management strategy

To manage the LifeWatch 'product', tactically and stategically, using the information contained within jira.lifewatch.eu.

Services, access, tools and facilities will be distributed among the participating countries

To become fully operational will take about 5 years. Therefore LifeWatch will:

- Adopt a phased approach, aiming for constant small increments in functionality with annual mandatory upgrades;
- Plan the first 3 years initially, taking input from User Platform and Data Resource Owner Platform;
- Create the roadmap; and,
- Undertake comprehensive review and re-planning as necessary at 2 year and 4 year points.

NOTE: The roadmap and quarterly development cycle will be presented in a similar manner to that shown for the Opencast Matterhorn project here: https://wiki.opencastproject.org/confluence/display/open/Roadmap

Product releases will be managed as 'agile' incremental releases on a quarterly cycle to grow the functionality constantly in small steps. This is what is most manageable from the point of view of functional integration. It also allows LifeWatch to respond reasonably quickly to new requirements coming from users.

There will be a process, that will need to be defined, for managing what goes into each release. Release content will be managed by the "Product Management Board" (PMB), with prioritisation of functionality based on input from the User Platform, the Data Resource Owner Platform, and the development team i.e., content of releases will be aligned to the interests and needs of Member States.

The PMB will strictly enforce technical quality and cut-offs so that release dates don't slip and the quality of the LifeWatch product is not compromised.

Product management aims and timeline:

- Year 1 (mid-2011 to mid-2012):
    - Foundations of Resource, Infrastructure and Composition layers
    - Provide support for some key Data Resource Owner categories
    - Provide some "Wow!" functionality in the User layer, based on several Showcases
    - Pursue R&D items on an 'innovation roadmap'
    - Release 1.1: Q2 2012
- Year 2 (mid-2012 to mid-2013):
    - Extend Resource, Infrastructure and Composition layers
    - Extend categories of Data Resource Owner to be supported
    - Add additional significant tools in the User layer
    - Introduce first elements of semantic mediation
    - Continue to pursue R&D items on innovation roadmap
    - Release 1.2: Q3 2012
    - Release 1.3: Q4 2012
    - Release 1.4: Q1 2013
- Year 3 (mid-2013 to mid-2014):
    - Iteratively build on foundations. Re-consider at 2 year review
    - Release 2.1: Q2 2013
    - etc.

The LifeWatch preparatory project is entering all construction details in an on-line database called "JIRA". This process facilitates updating and exporting parts of the information in other formats for stakeholder communities. At regular intervals new versions of the plan will be published on the LifeWatch web site www.lifewatch.eu

The JIRA database structure has three levels of detail:

- Projects (such as Application Services; Increased and targeted data generation; Service Centre; technical Construction
- Components (clusters of detailed issues within a project)
- Issues (Description, time line and budget to construct a specific issue).

5 Local Implementation, Acting locally

"Acting locally" means communicating with key audiences and engaging at the level of individuals, groups and organisations to bring about the emergence, added value and usage of the LifeWatch infrastructure.

The showcases are one means of doing that. Effective promotion of LifeWatch is another that leads to the identification of communities of scientists, such as the LagoonNET community, that can begin to see how LifeWatch may offer them a route to achieving something that they had previously felt to be difficult or impossible.

The following sections consider how to apply and interpret the global perspectives outlined above into actions that can be supported and performed at the local level.

5.1 Islands of infrastructure emerge and fuse

Subsidiarity[8] is the principle that LifeWatch will work by. “Islands” of LifeWatch infrastructure will be encouraged to emerge at the lowest organisational levels (interested research groups, interested institutions, interested regions, interested nations), wherever appropriate.

Over time, as individuals and groups cooperate and collaborate, we expect that such islands of infrastructure will organically grow and fuse (coalesce) with one another.

We will develop and promote a variety of supporting strategies to assist this process.

5.2 National infrastructure elements (national projects)

The national infrastructure elements are the elements that each country brings to LifeWatch i.e., data resources and specific tools and applications. They are linked to the priorities expressed by individual Member States, probably as reflected in their national projects. Descriptive information and status information about national projects is being collated in the National LifeWatch Networks space within the VE-Forum[9].

In part, LifeWatch overall construction strategy and the order in which we introduce functionality will be based on the interests of the funding Member States. The challenge for LifeWatch is to ensure that the national elements eventually link up with each other. This requires technical coordination and an interoperability plan.

Guidelines for coordination between LifeWatch and the National LifeWatch Networks are presently being formulated[10]. This is work in progress and is expected to evolve during 2010 as LifeWatch commences coordination with Sweden and other countries.

5.3 Citizen science, Observation Communities projects

Citizen science and amateur (or professional) observation networks represent an important category of new data sources for biodiversity.

There are already many such programmes running in most countries across Europe e.g., Sweden's ArtPortalen.

LifeWatch must have a supporting strategy to bring the data that these programmes collect into the LifeWatch framework and to assist with making it available. In part this is a technical issue but is governed by science policy.

---

[8] The principle that tasks should be devolved to and undertaken by the smallest, lowest or least centralised competent organisation.

[9] http://www.ve-forum.com/apps/comm.asp?$1=527

[10] Working document (Bánki, version 23rd December 2009) was sent out to National LifeWatch Coordinators at the end of 2009.

5.4 Supporting strategies

5.4.1 LifeWatch Supporting Project

This initiative allows projects to request a label as “LifeWatch Supporting Project” for projects that clearly contribute to the LifeWatch infrastructure development. The affiliated project has the right to use the LifeWatch logo in project communications and is valid for the agreed project duration. LifeWatch lists the affiliation on the main website.

Details of the scheme are available on the LifeWatch website[11].

5.4.2 Share-ownership scheme

The LifeWatch share-ownership scheme is the notion of motivating Data Resource Owners and Tool Owners to participate in and contribute to the LifeWatch infrastructure through their scientific interest, and to receive a “dividend” on their investment. Possible dividends could include:

- Wider use of data and tools, thus giving evidence (e.g., citations) and justification for seeking further funding for data collection or tool development activities;
- Quid pro quo – “if I publish my data I potentially gain access to your data”;
- Quick and easy route for data publishing;
- Gaining a level of influence over the activities and direction of LifeWatch evolution. Level of influence could be weighted according to the value or quantity of data contributed / used;
- Some kind of peer recognition e.g., concerning data quality. LCDP and LCTP marks at bronze, silver and gold levels. “Gold standard” providers;
- Quick(er) route to compliance with the requirements of the INSPIRE Directive;
- By agreement with national funding agencies in DP’s own country, could constitute fulfillment of funders’ data sharing / deposition policies.

5.4.3 LifeWatch Certified Data Resource

Should only be given to Data Resource Owners that meet a very strict set of criteria for the data they deliver.

A LifeWatch Certified Data Resource should be delivering 'gold standard' biodiversity data that is useful for a wide variety of purposes.

A LifeWatch Certified Data Resource should be able to demonstrate (by audit) the implementation of and compliance with a standardised workflow for the collection, accession, tracking, validation, annotation and tagging of their data.

NOTE: Here there is a need for the LifeWatch team to develop and coordinate an overall Action Plan for integrating known and useful biodiversity information resources.

5.4.4 LifeWatch Certified Tool

Similar to Certified Data Resource. See above.

NOTE: Here there is a need for the LifeWatch team to develop and coordinate an overall Action Plan for integrating known and useful biodiversity informatics tools.

5.4.5 LifeWatch National Coordinator (Champion)

This is somewhat like a GBIF node insofar as it acts as the national focal point for LifeWatch activities in the country. However, a National Coordinator or Champion is more likely to be an individual than an insti-

[11] http://www.lifewatch.eu/index.php?option=com_content&view=article&id=71&catid=1&Itemid=12

tution and should be a person willing to act as a promoter, champion, evangelist and catalyst for LifeWatch in their own country. The draft Guidelines for National LifeWatch Networks describes in more detail the role of a LifeWatch National Coordinator.

5.4.6 User Platform

The User Platform is a mechanism by which LifeWatch engages with its user community. It is the principal means by which LifeWatch informs and addresses its user community. The User Platform will be delivered through the LifeWatch Portal (see 4.11.5). A prototype[12] community / social networking website based on the HENVINET model has been produced by NHM London during Summer 2009. This will need to be brought within scope of the Portal activity and integrated at an early stage in the Construction Phase.

 See next section for some outstanding issues affecting UP that need to be addressed.

5.4.7 Data Resources Platform

The Data Resources Platform is a mechanism by which LifeWatch engages with its Data Resource Owners community. The Data Resources Platform will be delivered through the LifeWatch Portal (see 4.11.5).

 Several outstanding issues need to be addressed:

- A key attraction for Data Resource Owners in publishing their data to LifeWatch ought to be the possibility of gaining better connection to those people using their data. It is not clear how the current model supports this and it seems that this has not been addressed adequately by WP3 and WP7.
- A stronger case around the benefits of contributing data to LifeWatch needs to be developed and communicated to DROs (both organisations holding data and individuals holding data).
- It appears that DROs may be concerned about who will be using the data and about the 'rights' associated with that, rather than looking first at the opportunities enabled by LifeWatch and then addressing the rights question. Is the concern that the NoEs are afraid that the LifeWatch ERIC will start commercial business on top of the data? If so, this needs to be addressed in the financial flows.
- A current concern is the apparent separation between the DRP and the UP.

5.5 Engage at the level of the individual

Engaging with individuals is likely to be the most effective way of encouraging take-up of LifeWatch services and facilities. This will be true regardless of whether the individual is a scientist or policy-maker wishing to use data and tools offered by LifeWatch, or an individual with data and/or tools to share with the LifeWatch community.

Overcoming barriers to engagement at the level of the individual is the single biggest challenge facing LifeWatch once it enters its construction and operation phases. There are several barriers standing in the way of this, including:

1. Failure to adequately address an individual's scientific need, operationally and/or technically;
2. Failure to clearly explain and match the benefits of LifeWatch to an individual's need;
3. Making it too complicated for an individual to engage (for example: learning curve too steep, inadequate return on an individuals' investment of time, failure of flimsy methods and procedures, etc.).

[12] http://lifewatch.nhm.ac.uk/drupal/lifewatch/

We will need to devise some clever strategies to overcome these barriers. But, we can only do this against a background of clear psycho-sociological understanding about what it is that makes individuals and organisations commit their time, effort and other resources to an enterprise like LifeWatch i.e., we must more clearly understand the psychology of motivation and engagement.

5.6 Strategies to engage the individual

LifeWatch has to position itself to differentiate clearly from the multitude of other on-line initiatives that are taking place. This is clear from the responses to the WP7 survey.

We should not position LifeWatch on the 'e-Science' ticket because that vision is too far removed from what the general community is familiar with today. They don't / won't get it. Therefore, even positioning has to take a phased approach, gradually introducing more elements and building up the vision over a period of time.

Engaging at the level of the individual means making LifeWatch relevant to the everyday needs of that individual. It means:

- Raising awareness at the local level, using examples relevant to that individual's specific context, to demonstrate "what's in it for me?";
- Keeping tools and methods simple;
- Using, in addition to english as a common language, the local language where that is appropriate;
- Matching the offered data and tools to the role of the individual (research, policy, etc.); and,
- Offering local training and day-to-day support.

Clearly, the potential user community is extensive and cannot be tackled en-mass. Engagement must be prioritised and themed. Themes must be related to ecosystem services and address specific societal issues of local importance.

5.6.1 Personalisation

Engaging at the level of the individual means providing the ability to personalise LifeWatch to the needs of that individual. This can be achieved by configuration options within the LifeWatch Portal. e.g., an individual must be able to impose their own preferred taxonomic model on taxonomic data. They must be able to restrict views to tools and data relevant to their sub-domain of interest, if so desired. They must be able to display chosen gadgets/widgets/portlets on their personal Portal page, etc.

5.6.2 Social networking

Social Networking is a key strategy for engaging individuals (cf. FaceBook, LinkedIn, Nature Network, etc.). However, the number of scientists presently making use of social networks is still quite low. Thus, strategies will be needed in which membership and participation in a social network provides opportunities that would not be accessible by other means e.g., the possibility to discover potential collaborators, the possibility to access specific funding initiatives, etc.

We need to identify a social networking solution, perhaps developed specifically for the scientific community, that we can tap into. We don't want to have to re-invent this or to duplicate what scientists may already be using. To what extent, for example, could we integrate "Nature Network" (http://network.nature.com/)? This is wider research infrastructure issue that should be taken up during the Construction Phase. Ideally, we would like to see something like Nature Network become as essential to our research community as some of the Thomson Reuters products are to the academic community generally. A bigger issue than LifeWatch but we can lead the way maybe?

myExperiment is another example of scientific social networking capability that may be extensible to meet specific LifeWatch needs. However, we should consider perhaps, how to avoid establishing yet another proprietary social neworking tool. A first step could be for a small community to instantiate and configure / customise myExperiment and its companion tool BioCatalogue, for the biodiversity realm.

5.6.3 First steps to data discovery and visualization

The immediate requirement at an early stage in LifeWatch capability is the functionality to support data discovery and the discovery of associated / linked data.

The approach described in this book chapter on "Data Finds Data" is one method of doing it.

Once some data has been discovered, what is the next thing that the scientist will want to do with it? Visualize it perhaps?

5.6.4 Encourage sub-communities in bottom-up efforts

Consider the model as a number of small close knit researcher communities with frequent interactions between individuals, existing within a universe of more remote communities with which less frequent interactions occur.

Encouraging small sub-communities to work closely together leads to bottom-up emergence of results - data, tools, infrastructure, analysis, science. Typically, the interactions between the individuals within such a community are frequent. Also, typically, there will be less frequent interactions between such communities that can be made more effective by an overlay of top-down, global guidelines that have the effect of fusing the sub-communities together.

Such global guidelines have to be packaged up, delivered and made very easily accessible into the work of the sub-communities. The sub-communities must contribute to creation of the global guidelines in order to achieve the buy-in.

Strategies to use, for example, include:

- Encourage the publication of data and provide a mechanism for giving credit for the re-use of that data;
- Encourage the sharing of workflows and provide a mechanism for allowing modification (improvement) of those workflows by others;
- Preserve / curate the results of such community activities. Make them discoverable by others. Make it possible for others to join on to add extra value;
- Advertise information about relevant tools widely;
- Provide mechanism(s) by which individuals / small communities can find other like-minded individuals and communities.

5.6.5 Support creation of semantically enhanced publications

It must be easy to for users to create semantically enhanced publications, in line with emerging trends in scientific publishing. There are examples of where this has been done already, both in the biodiversity domain and elsewhere. The approach of NLM/Taxpub deals with just one aspect.

LifeWatch must support the complete chain of scientific process in this regard: i.e., collect/discover data; analyse data; publish results, with links to preceding elements.

However, this is an area where other disciplines and the publishers themselves are innovating too, so LifeWatch must not proceed in isolation.

We should consider a construction category for this area, also adding semantically enhanced publications to the list of differentiating areas for LifeWatch.

5.6.6 Create an environment to support Open Notebook Science

Open Notebook Science (ONS) or Open Science or Open Research is the idea that science should be conducted in an open and public manner through frequent / continuous publication of method, data re-

sults and conclusions on the internet (in blogs, on websites, etc.) and the encouragement of discourse about those aspects.

What started in the Chemistry community could become a powerful force in the biodiversity informatics community when combined with the ideas of data publication and discovery promoted by GBIF, the use and sharing of workflows as promoted by myExperiment, the tracking of provenance and data re-use/citation (Open Provenance Model), and a semantic mediation framework.

We could do this by providing a customised LifeWatch-wide wiki, either maintained by LifeWatch on behalf of the community (e.g., Atlassian's Confluence product) or outsourced to a provider like Wikispaces (for example).

NOTE: We have set up an instance of Atlassian Confluence[13] for use by the LifeWatch team. We could make this more widely available.

5.6.7 Linking LifeWatch-RM to the individual

Linking the LifeWatch-RM at the level of the individual scientist or at the level of a specific research group is perhaps the most important mechanism for LifeWatch to get right[14].

The LifeWatch-RM is complicated and expressed in ICT jargon, making it difficult for non-ICT specialists to comprehend. An approach to LifeWatch compliance based on top-down imposition of the LifeWatch-RM guiding principles will fail for this and several other sociological reasons.

We can think therefore about separating elements of the LifeWatch infrastructure into those where it is essential that the principles of the RM are applied and those where that is not so essential. In fact, we can imagine a gradient, outwards from the core systems of the LifeWatch infrastructure. Sub-systems at the centre are essential for LifeWatch operations (= core ICT infrastructure, 4.11). For these, there is a strong requirement for compliance with the RM. As one moves progressively towards the edge of the infrastructure, the requirement for less essential sub-systems to comply with the guiding principles becomes weaker. One can see that various different LifeWatch Data Resource Owners and Tool Owners will appear at different points along that gradient. Those closer to the edge will have an easier time to integrate their work with LifeWatch, whereas those closer to the centre will have to work more closely with and understand in more depth the complexities of the overall architecture.

We can represent this as a series of concentric rings.

<draw an appropriate diagram that shows this>

At this stage, 5 "rings" can be imagined:

1. Core resource: Mandatory compliance with RM principles is necessary.
2. Essential resource: Compliance with RM principles to the fullest extent possible is highly desirably. Interoperates well with other LifeWatch resources
3. Desirable resource: .....
4. Minimum compliance resource: All that is necessary to register a minimum compliance data or service resource with LifeWatch is a description of the data or service resource. Interoperability is limited.
5. Beyond LifeWatch - no compliance with RM principles is required. Ad-hoc interoperability may be achievable without involvement of LifeWatch

13 https://wiki.lifewatch.eu/

14 A discussion with Axel Poigné at the end of July 2009 established the seeds for how to do this and forms the basis of the description in this sub-section.

5.7 Make it easy to create LifeWatch applications

LifeWatch applications are LSNs (LifeWatch Service Networks, 4.10.2) that are comprised of service instances selected from the LifeWatch Base Services and from appropriate thematic service groups.

The procedures for creating LifeWatch applications compliant with the LifeWatch-RM rules must be simple. LifeWatch applications can be easily created by one of a number of methods:

- Workflow composition tools - Use a workflow tool, such as Taverna, Triana or Kepler, to compose workflows made up of resources and service instances "discovered" and made known to the tool. The tool can discover (or be told about) the various services in the global Base Service and Thematic Service groups and also about the various data resources, again organised by category;
- Sharing of workflows – use workflows already created by others and shared, for example, through myExperiment;
- Integrate existing tools.

NOTE: Here there is a need for the LifeWatch team to develop and coordinate an overall Action Plan for integrating known and useful biodiversity informatics tools.

5.8 Routes to integrating existing tools

See also share-ownership scheme, 5.4.2 above.

Integrating existing biodiversity informatics tools is an important mechanism by which LifeWatch can increase its functionality. Such tools have been developed over periods of, perhaps, several years and may already have existing user communities.

There are several degrees of tool integration possible. The degree of integration to be achieved is mainly to be determined by the desires of the tool owner / provider in consultation with LifeWatch. Degrees of integration, proceeding from loose to tight, can be considered as follows:

1. Access to the tool for LifeWatch users is provided as it presently stands e.g., as a separately downloadable software package, via an existing web portal URL/interface, etc. LifeWatch provides a link / entry point to the tool from the LifeWatch portal / environment. No support is provided by LifeWatch.
2. The tool may acquire a basic element of LifeWatch branding, such as "LifeWatch Certified Tool".
3. The tool is partially integrated into the LifeWatch environment. Direct access to tool functionalities are provided by the tool owner / provider through an existing URL/portal interface. A service-oriented interface to the tool functionality is provided to the LifeWatch environment such that the functionalities can be discovered and used as toolbox items in the workflow composition environment.
4. The tool may acquire a basic element of LifeWatch branding, such as "LifeWatch Certified Tool", or higher degree of LifeWatch 'look and feel' branding.
5. The tool is fully integrated into the LifeWatch environment and re-branded as a LifeWatch tool. Tool functionalities are available as LifeWatch services, delivered to users as portlets in the LifeWatch portal and toolbox items in the workflow composition environment. Support and further development of the tool is undertaken by LifeWatch own-staff or by the original tool-provider under negotiated contractual arrangement with LifeWatch.

Achievement of one of the looser levels of integration does not preclude tighter integration at a later date.

As an example, one may take the case of the EDIT Platform Cybergate. Cybergate presently provides (or shortly will provide) access to a wide range of capabilities for taxonomists. It is currently available at the URL: http://dev.e-taxonomy.eu/platform/. The Cybergate re-directs the user to his/her choice of tool, again at specific URLs. Functions of some of the tools are presented as web services, for example: CDM REST services, MAP service, etc. Partial integration of the tools to LifeWatch could involve publishing the service details in the LifeWatch Service Registry, whilst leaving the existing EDIT branding on the present URLs. At a later date, the existing EDIT branding could be replaced with LifeWatch branding, accompanied by a move of URLs to the LifeWatch domain.

Routes for tool integration need to be specified and appropriate procedures put in place. See 4.10.4 above.

An interesting survey of tools, undertaken by the Atlas of Living Australia, can be found here: http://alatools.pbworks.com/ (or PDF). The format it uses for describing each tool is interesting to consider as a basis for describing tools available through LifeWatch. Such a set of descriptions could be maintained by and accessed through the Service Centre.

Mgmt Action Needed: Specify tools and order of integration, broken down by Release, in a manner that connects to each major research area to be supported by LifeWatch. Probably, by themes.

5.9 Making existing data available

Making existing data available through LifeWatch has a cost associated with it. Therefore, Data Resource Owners have to be shown the benefits that come from publishing. Some DROs will already be convinced or open-minded about this and these are the ones that we can engage with most easily.

Benefits of making data available include:

- Create of new opportunities for data use.
- Promotion of value: Data publication increases publisher visibility so long as the identity of the data source is preserved throughout subsequent (re-usage). Tools that track and show data re-usage can demonstrate value.
- Maintainenance of interest: Connect publishers to users of their data. Provide quality feedback mechanisms. Provide tools for quality improvement. Provide tools to ask interesting questions.
- The culture is changing to favour funding of those DROs that do publish. <insert explanation here of the triangular legal arrangements LifeWatch will adopt to encourage data publication>.

DROs that have expressed an interest to engage during the first phase are:

- Centre for Ecology and Hydrology (CEH), Lancaster
- Netherlands Institute of Ecology (NIOO)
- Muséum National d'Histoire Naturelle (MNHN), Paris

Several other organisations (e.g.., from Hungary, Finland, Belgium, Sweden, etc.) have also shown interest but further qualification is needed before these can be allocated to particular releases. The processes of creating the data catalogue and determining the first applications should help with this.

See also share-ownership scheme, 5.4.2 above.

5.9.1 Create a data catalogue

Creation of a catalogue of the different categories of data and of the data providers and datasets to be offered to LifeWatch has been seen as a fundamental requirement of the technical planning.

We have some knowledge of the kinds of data and the likely Data Resource Owners but no definitive list around which to plan. This issue was discussed at the 5th MC meeting in Brussels, June 2009.

The MC agreed that a template for data catalogue entries, based on the metadata requirements of the INSPIRE Directive should be created and that progress should be made, probably starting at the 2nd Data Resource Owners' Workshop, towards compiling specific entries for such a catalogue.

It is perhaps desirable that the structure of such a catalogue should be compatible with the structure of the emerging GBIF Global Biodiversity Resource Discovery Service. This needs to be further investigated.

Considering that collecting information about all datasets to be put on-line in LifeWatch when it starts operations is a major exercise, some constraints are needed. In the first instance, we will limit this to col-

lecting metadata about the datasets that are needed on-line in LifeWatch when it first commences operations i.e., those datasets needed to support pilot projects and some of the first stage (release 1 services).

The datasets that will be brought on-line first will be determined by the needs of the first research areas and applications.

Are the first applications in effect, the national projects? They are not the showcases because, by definition, the showcases are only being used to derive the Biodiversity Research Capabilities to be supported (which will be represented in the Base Services and the Thematic Services).

5.9.2 Categories of data

Categories of existing data that LifeWatch can accommodate include:

- Specimen data e.g., from Natural History Collections. Can we rely on GBIF to provide all of this?
- Terrestrial (incl. freshwater) ecological data
- Marine ecological data
- Taxonomic data
    - PESI provides part of the taxonomic foundation for LifeWatch. PESI delivers the standardised, that is, the most authoritative and most complete species list for Europe (for multicellular life at least). It is a reference list, much like a gazetteer. This list is (will be) further linked to the Global Names Architecture and to one or more taxonomic (phylogenetic or management) classifications, such as the Catalogue of Life or the CBOL classification. This list is also linked to the species names in the nomenclators (ZooBank, IPNI, etc.), both directly, and indirectly through the GNA / GNUB.
    - The Catalogue of Life provides authoritative data about species names, common names, synonyms, taxon concepts and a particular taxonomic hierarchy.
- Genomic data e.g, sequence data, DNA barcodes, etc.
- Remote sensing data i.e., coming from satellites and radar

5.9.3 Wrapper development kit

At this moment it isn't clear whether a single wrapper development kit would suffice to support both these (and other cases) or whether separate kits more tailored to specific communities of use would be advantageous

A wrapper development kit (WDK) is a collection of items (documentation, design patterns, source code examples) that make it easier for a Data Resource Owner to connect their datasets to an infrastructure. A WDK helps DPs to keep their costs down whilst at the same time allowing the LifeWatch team to retain some control over the standards for data, metadata and protocol operating at the interface of the data source and the infrastructure.

LifeWatch WDK should support relevant TDWG standards and should ensure that in so doing, the data delivered by DPs to the LifeWatch infrastructure maintains its integrity i.e., should not be altered in any way. Supporting the relevant standards (TCS and TAPIR?) keeps costs down.

The GBIF IPT seems to be one kind of WDK supporting GBIF's requirements for accessing 3 types of data: taxon primary occurrence data, taxon checklists and general resource metadata.

5.9.4 Procedures for "admission" of a Data Resource Owner

See Data Resource admission process (4.10.3)

Including details about licensing agreements. For example, there are 4 possible licensing arrangements to choose from:

1. Directly between providers and users;

2. Between providers and users having LifeWatch as intermediary
3. Between providers and Life Watch
4. Between LifeWatch and users

5.9.5 ORCHESTRA approach to source system integration

ORCHESTRA Service Network (OSN) tutorials provide some hints about practical approaches to integration of external source systems. We should develop a set of guidelines based on these.

5.9.6 GBIF support for making resources available

There are a number of activities underway within GBIF that are concerned with making external resources available. IPT is one example. LifeWatch should avoid duplicating such activity and should work closely with GBIF to ensure that the sector-wide tools that emerge are sufficiently general purpose to support both GBIF and LifeWatch needs.

5.9.7 Access to data and tools "free at the point of use"

Access to data and tools should be "free to users at the point of use". This does not mean that access to all tools and all data is free-of-charge. There must be mechanisms in place for remunerating Data Resource Owners and Tool Owners appropriately.

This will be one of the biggest challenges for LifeWatch.

An instructive model is EDINA. EDINA is the UK's national academic data centre, with the remit to provide on-line data services to support the UK academic community. EDINA services are diverse and wide-ranging; covering article references journal catalogues, ebooks, maps and geographical data sets, agricultural data, film, image and sound archives, and much more. EDINA services are available 24 hours a day, 7 days a week. They are free of charge at the point of use. Access to most services is enabled by institutional subscription, followed by authenticated access from that institution's end-users. This is regardless of whether those users are working "on-campus" or remotely. In the latter case, an authentication mechanism such as ATHENS (deprecated) or Shibboleth is used.

For LifeWatch to adopt such a model will would require developing sufficient "added-value" to motivate institutional subscription.

How would this be compatible with the envisaged national funding model for LifeWatch?

Another possibility to consider is “pay-to-publish data”, in which the data provider gains a quick, easy route to data publication (e.g., based on the Scratchpads model) where all the hard work is already done. As with pay-to-publish journals, there could be an element of peer review leading to the granting of a quality mark for the data.

5.9.8 Action Plan for the digitisation of NHC specimens

The LifeWatch strategy includes plans to create an effective system to enhance access to the information contained in Europe's natural history collections. European collections are special because of their rich historical holdings, and the large amount of specimens from outside Europe, especially tropical countries.

GBIF informs users about the available digital data, but only a very small fraction of specimens have been digitised to date, perhaps 2%. In the collections, access to physical specimens is optimised for taxonomists: the arrangement follows species names and organism groups. Access to information about where the speciment was collected, when and by whom is thus inefficient for non-digitised collections. The optimal solution would be to digitise all specimen holdings but because of the high costs involved (2-30 $US per specimen for quality data) this is impossible. Generally, collection institutions are willing to make records freely available but lack the resources and funding opportunities to do so.

User demand must therefore drive the costly digitisation of individual specimens. To make this possible, collections have to be enabled to answer the questions, with a clear priority on WHERE (geographic location). Example: "How many undigitised specimens of Amphibians from Tanzania exist in which collections? How much does it cost to digitise them?"

Once this information is available, biodiversity researchers and countries interested in their biodiversity can set the priorities for the digitisation process.

The investment needed to create the needed metadata catalogue, although much lower than in the "digitise it all" approach, is considerable.

Some preliminary work has already been done but there is significant further work to do to work out the best ways to assemble the necessary metadata, optimised for appropriate taxonomic and geographic categories.

This further work needs to be planned and costed.

5.10 Connecting with citizen science and observation community data

This category of data should be treated separately due to the significant number of separate and diverse initiatives in this area.

5.10.1 Wrapper development kit

At this moment it isn't clear whether a single wrapper development kit would suffice to support both these (and other cases) or whether separate kits more tailored to specific communities of use would be advantageous

5.11 Connecting sensors networks

<something needs to be said here, perhaps coming from the BioDivGrid workshop outputs?>