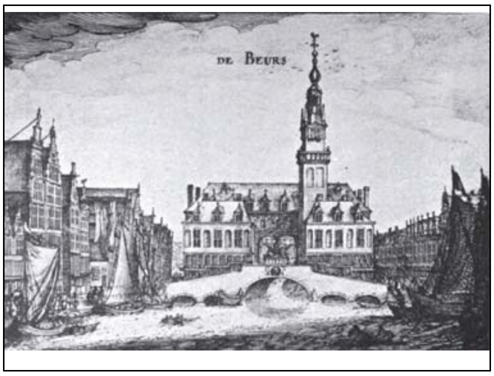
LifeWatch – An e-Science infrastructure for biodiversity research

A keynote talk to the EGI Technical Forum, Wednesday 15th September 2010, Amsterdam by Alex Hardisty, Director of Informatics Projects, Cardiff School of Computer Science & Informatics, Cardiff University.

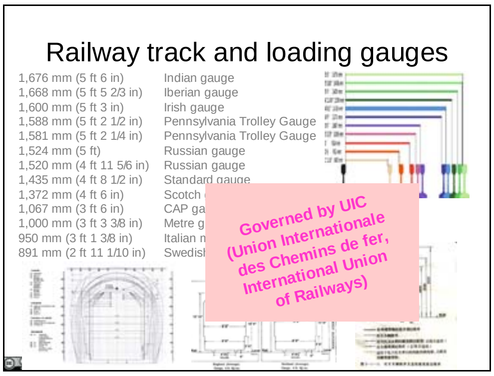
Slide 1,2,3

Hello. Good morning.

Like me, most of you will have travelled some distance to be here today. I hope that you didn't have to use a ship or a horse to get here.

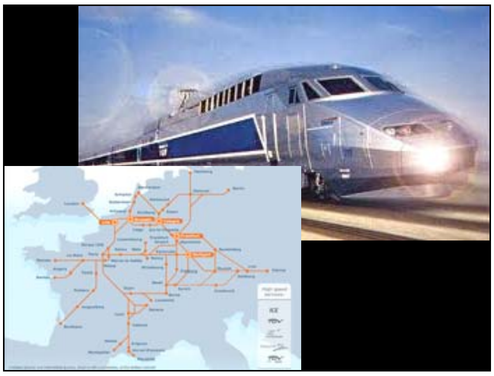


I hope that you were able to take advantage of a more integrated transport system – plane, train and tram. A transport system that works across national borders and regional boundaries, a transport system that is offered to you by multiple transport companies with, in the case of railways for example, a standard track width and loading gauge with coherent signalling, timetabling and ticketing systems.



Slide 4

If you came by high-speed train – ICE train, Thalys, or TGV, you will have found this to be so, although not if you came from the Iberian peninsula, Scandanavia or the Baltic states. Increasingly, Europe's high speed railways are standardising around a single track and loading gauge and signalling and safety systems that will make train travel from one part of Europe to another more of a pleasure than a trial. Behind the scenes this is the result of regulation, harmonisation,

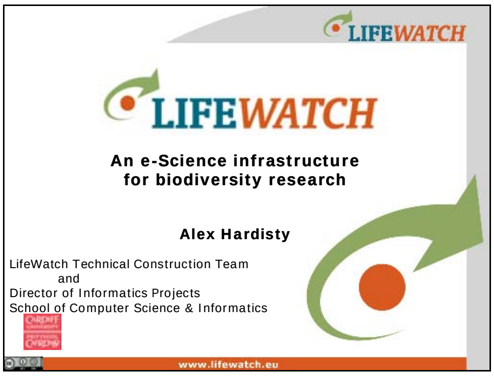


standardisation and lots of hard work by a great many people.

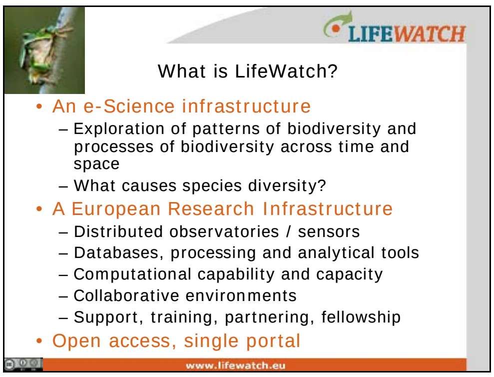
Similarly, this is the biggest challenge the ESFRI research infrastructures face during their construction phase – the socio-technical challenge of bringing communities together and uniting them behind common technical approaches that span institutional boundaries and cross international borders.

For EGI's role in this, this work has already started
Slide 5,6 under the leadership of Bob Jones in the European e-Infrastructure Forum … but the individual ESFRI infrastructures have their own roles to play too.

This morning I want to tell you about one of those infrastructures, LifeWatch and the particular challenges it faces in trying to deliver ICT infrastructure for a community of 25,000 plus end-users.

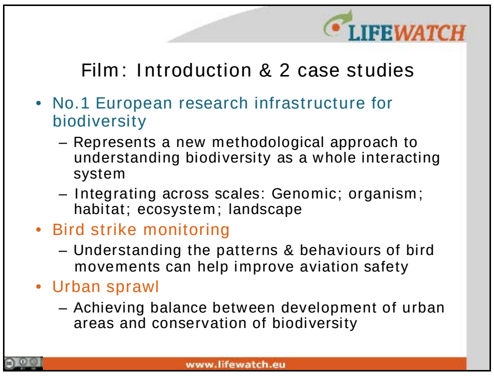


I'm a member of the LifeWatch Technical Construction team. With my colleagues I'm responsible for the technical strategy for LifeWatch and for developing a blueprint for constructing its ICT infrastructure.



LifeWatch is an e-Science infrastructure for biodiversity
Slide 7,8 research. The science of biodiversity is the study of the diversity of plant and animal life on our planet and the environments they live in. LifeWatch will allow scientists to explore, describe and understand patterns of biodiversity and the processes that maintain it in space and time. It will help scientists to answer the question "what causes species diversity?" and to understand how man's impact on our environment affects that diversity.

You may not have heard it yet, but in this International Year of Biodiversity, the UN Environment Programme and 86 governments have requested the UN General Assembly to approve the creation of an Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES). With concerns about conservation of biodiversity, climate change, food security, and human health and well-being high on the political agenda, this is clearly an important focus for future attention and investment.



Slide 9

I want to introduce LifeWatch to you with a 10 minute video that, with the help of 2 case studies, shows what it is intended to do.

Bird strike monitoring illustrates what can be achieved by integrating information from multiple sources in a 'system of systems'.

Urban sprawl illustrates the importance of being able to discover and access necessary data across organisational boundaries and highlights the need for LifeWatch to support collaboration between users and across datasets – a requirement echoed in the soon to be published report of the High Level Experts Group on Scientific Data.

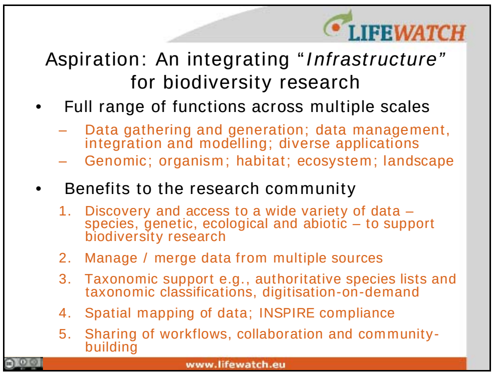
The case studies give an insight into both the potential societal benefits of a research infrastructure such as LifeWatch and also into the complexity of its distributed data and computing needs. Like transport networks, LifeWatch aims to be a broadly based research infrastructure and these two case studies barely scratch the surface of the many scientific uses to which the infrastructure will be put.

SHOW VIDEO: http://www.youtube.com/watch?v=hSEGW1slYNg

Slide 10

The 'Unique Selling Point' of LifeWatch is that it is an *infrastructure*. That is to say: It's the permanent elements that are needed to create an internet and web-based system that links personal and institutional systems - data resources and analytical tools - for biodiversity research. Human resources provide appropriate support and assistance to users.

Slide 11



In this sense, the aspirations for LifeWatch are to provide a full range of functions across multiple scales for: data gathering and generation; data management, integration, and modelling to support diverse applications. These functions will enable discovery and access to a wide variety of data - genetic, ecological and environmental - to support biodiversity research and policy.

LifeWatch will support globally unique identifiers for biodiversity resources. This is both for physical assets (particular datasets, for example) and biodiversity concepts (for example, a species concept or an ecosystem definition). As a result, managing, merging and manipulating data from multiple sources will be much easier than it is at present. Unique identification of concepts aids clearer understanding and helps to resolve ambiguity.

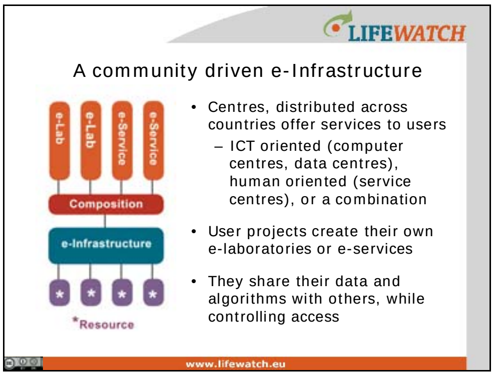
LifeWatch will support workflows, and sharing of workflows, as the paradigm for accomplishing specific research tasks that involve transformation, processing and analysis of data. This will lead to better collaboration and community building.

It will support mechanisms of provenance to permit tracing of data and workflows for reproducibility of scientific analysis, and tracing of data re-use and citation.

It will support spatial mapping and the requirements of the INSPIRE Directive on the availability of publically held spatial information.

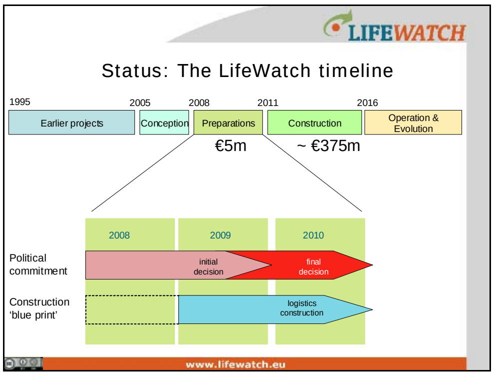
Finally, over the long-term, LifeWatch aims to support semantic interoperation of heterogeneous data and tool resources.

Our deployment approach is a distributed approach,
Slide 12 with data, tools, services and facilities being distributed among participating countries. Our deployment revolves around the idea of *centres* offering *services* to users. Such centres can offer services that have a predominantly ICT focus – for example, data repositories and computational model capabilities, or they can be predominantly human oriented. That is to say, they can offer human services like a helpdesk or a partner matching service. Combinations are also possible, and expected.



Driven by scientific need, LifeWatch takes its requirements input from the communities of users and the data providers that it serves.

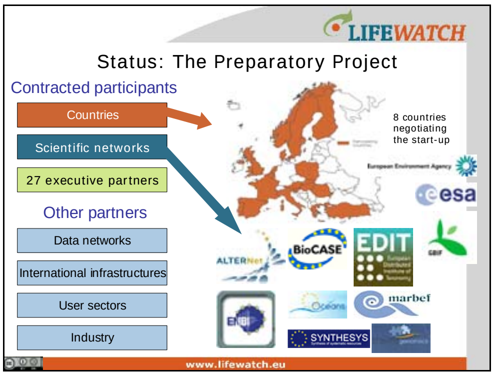
In this respect we are secure. Various projects
Slide 13 organised over the past fifteen years have already tackled components of the problem. In 2005 large communities of researchers sat together and conceived the scientific case for LifeWatch. The European Strategy Forum on Research Infrastructures accepted the scientific case and LifeWatch was included in the ESFRI roadmap. With Framework Programme 7 funding the Preparatory Phase was started in 2008.



During the preparatory phase we have two main tasks: firstly, to secure commitment from as many countries and stakeholders as possible to join LifeWatch, and secondly, to prepare the blueprint for the construction phase - not only technical, but also legal, financial, and organisational. Construction will be a significant undertaking. To become fully operational will take 5 years and we expect it to cost in the region of 375 million euros.

Only a proportion of that is likely to be "new money".

The emphasis of the financial model is on aligning already budgeted national expenditures for biodiversity research infrastructures with the aims and objectives of LifeWatch. In any case, we adopt a phased approach that will align incremental increase in expenditure to constant small increments of released functionality.



Slide 14 The preparatory phase project is being conducted by 27 executive partners, of which my own institution, Cardiff University, is one. These partners represent the countries and the scientific networks and other organisations that have an interest in LifeWatch.

19 country governments have given their support to the preparations, with more than half of those having prioritised LifeWatch as an important infrastructure on national roadmaps for research facilities. Last week 8 of those countries began detailed negotiations on the establishment of the LifeWatch legal entity and the financial commitments to support it.

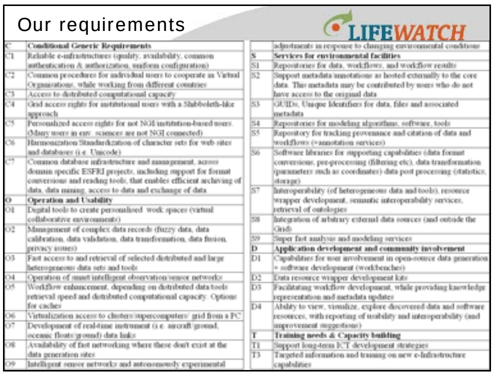
The scientific networks of excellence are the founding fathers of LifeWatch. There are many more networks that are interested in and waiting for LifeWatch, but these are the ones that have worked continuously over the last years to bring the LifeWatch vision to fruition on behalf of their member institutes. Each network has around 40 to 60 member institutes, so these alone already represent some 5000 scientists.

ALTERNet deals with Terrestrial ecology. BioCASE EDIT and Synthesys cover taxonomy. ENBI is the European Network for Biodiversity Information. Eur-Oceans and MarBEF (now MARS) cover marine ecology and MarineGenomicsEurope joins various communities to ensure Europe’s lead in marine genomics is maintained.

Apart from these executive participants, we are in discussion with other networks and organisations: non-Governmental organisations like GBIF, the Global Biodiversity Information Facility, with whom a Memorandum of Cooperation has recently been signed; the European Environment Agency; the European Space Agency, and others.

Industry is not only interested as a potential user of biodiversity data and tools, but also as a co-developer of the infrastructure and we are in discussion with two major IT vendors at the present time about the software development and deployment model for LifeWatch.

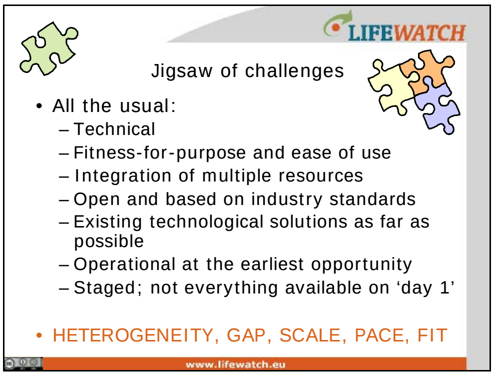
Slide 15



As you would expect, we have a wide range of detailed technical requirements. Actually, many of these requirements are not specific to the biodiversity domain. They are common across a wide range of environmental science infrastructures. In LifeWatch we have specific nuances of those requirements. We have our specific formats for recording data. We have specific vocabularies for biodiversity terminology, concepts and objects. We have our specific experimental, observational and recording methods, ranging from the small-scale and manual – specimens, metre quadrats thrown on the ground, to automated and large-scale sensor networks and colour-enhanced satellite imagery.

Slide 16

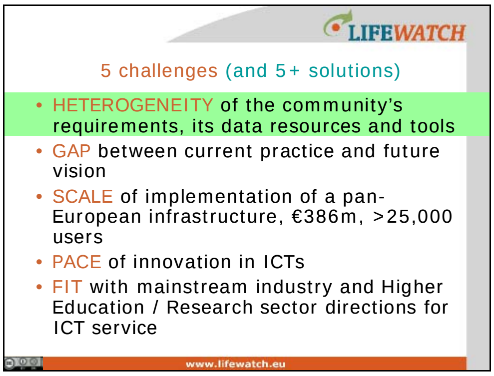
These and other requirements present the usual jigsaw of challenges you may expect to be associated with engineering a large-scale complex state-of-the-art IT infrastructure. There are technical challenges associated with meeting specific requirements. There is the issue of fitness for purpose and ease of use, etc. etc.



However, there are five key challenges that I want to highlight. They are what I call: Heterogeneity, Gap, Scale, Pace and Fit. These don't come from the requirements themselves but from the characteristics of the context in which we find ourselves – the construction and operating constraints. That is to say, they are challenges that influence our thinking about how we will achieve the requirements of LifeWatch. I want to focus on them because they shape the technical approach that we have based our construction plans on. And I think they are relevant when it comes to thinking about ESFRI infrastructures in general and how EGI can support the diverse needs of a wide variety of different communities.

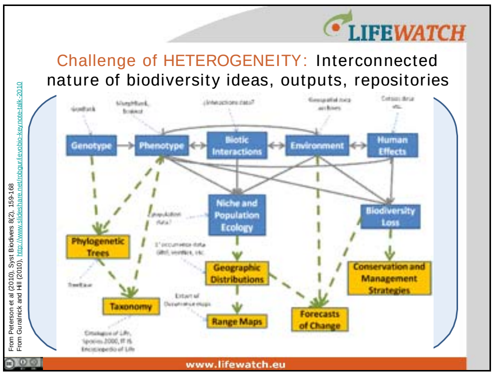
Slide 17

Firstly, there is the heterogeneity of the biodiversity community itself, its requirements, its data resources and its tools.



Slide 18

By its nature biodiversity science spans a number of more familiar disciplines: biology, botany, zoology, ecology, genetics, soil science, biogeography, climate science, chemistry - to name but a few. Each of these established communities already has its own way of doing things, their own data and their own tools. Not only that, but they have their own different vocabularies and conceptual underpinnings.

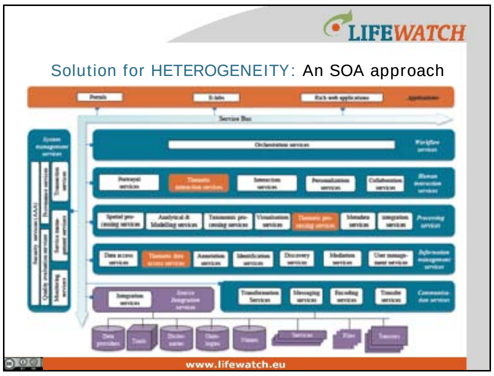


This picture illustrates the intricacies of the interconnections between biodiversity ideas (blue), sources (black text), and outputs (yellow). GenBank, environmental data and census data may be familiar to you. Geospatial data archives is an innocent term that hides who knows what! Others sources, such as primary occurrence data are less well known but equally extensive. GenBank holds more than 100 million records. GBIF gives access to data about holdings in natural history collections – more than 203 million records mobilised and increasing by 20% per year.

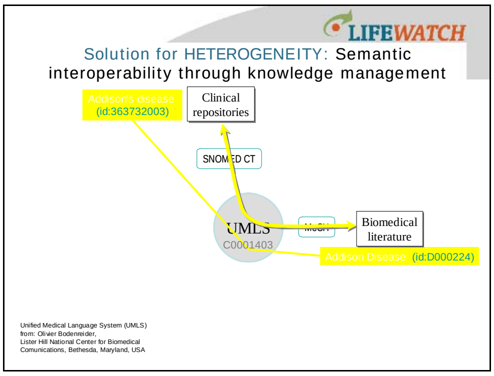
There are numerous smaller datasets in the hands of individual researchers. If computerised at all, these are often held in spreadsheets and probably with no identifiable common structure. They are not shown in this picture but there are thousands of them.

The challenge of heterogeneity can be addressed by several technical mechanisms. I will mention two. But before technical solutions, we also need what I call people solutions. By that I mean that we need the involvement of enlightened scientists, expert in their own domain but also enthusiastic, trained and technically competent with the new technologies of e-Science, working alongside engineers. The Service Centre function of LifeWatch is intended to support this – to help users to explore, learn and put into practice a new modus operandi for biodiversity research.

Slide 19 For our first technical solution we adopt a service-oriented approach. By wrapping resources and tools as services with interfaces and with metadata, we are able to hide away the heterogeneity of the underlying implementations. We can provide a common means for integrating data resources and modelling capabilities into the infrastructure. We can provide a set of guidelines – a cookbook of procedures – for how to do that. We can also provide common solutions to problems of discovery, data transformation, messaging and event notification, management and monitoring, and security.

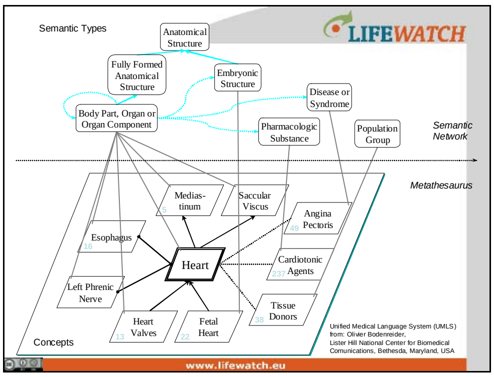


Secondly, we are exploring a knowledge management
Slide approach from the field of medical information systems
20,21 to see whether it can help us to resolve differing vocabularies and conceptual understandings that exist in different parts of our community and to create computer applications that more intelligently comprehend that language. In this and the next slide I use medical examples because the equivalent work in the biodiversity field hasn't been done yet.



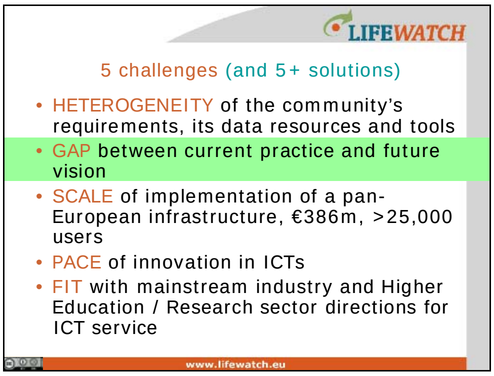
The approach, called the Unified Medical Language System or UMLS is a system pioneered by the National Library of Medicine in the United States. UMLS integrates and distributes key terminology, classification and coding standards in the field of medicine. Its components underpin a wide variety of applications in that domain, of which perhaps the best known here will be the PubMed citations database for biomedical literature.

Put simply, UMLS works in two parts. Firstly, by creating a metathesaurus it links concepts from one vocabulary with those of another – Addison's Disease in this example – and second,
using the heart as the example, it consists of an overarching semantic model that links key concepts in the metathesaurus one with another in a commonly accepted domain framework, for example about anatomy.

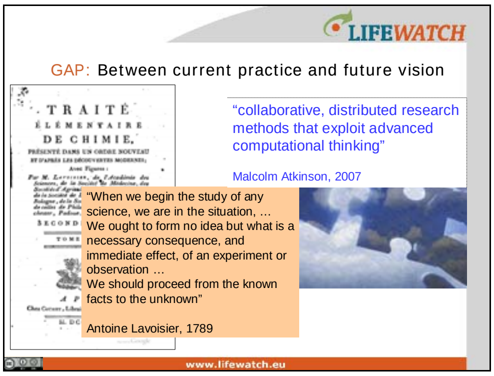


Slide 22,23,24

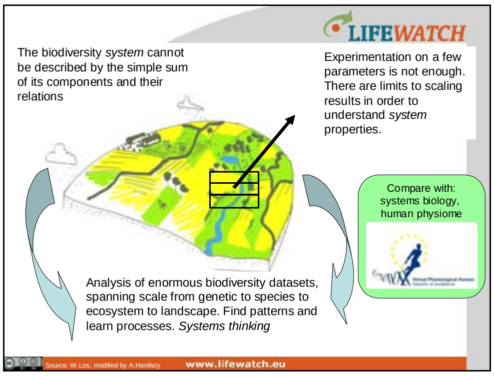
Something that has become apparent to me as I have worked in e-Science generally and on the LifeWatch preparatory project in particular is the extent of the gap between the current everyday practices of scientists working today and the blue sky vision of a future for collaborative in-silico science that e-Science heralds.



To the initiated, and that probably includes a great number of people in the room today, to those who grew up with early e-Science projects like the LCG, GridLab, myGrid, and AccessGrid, the advantages and benefits of taking an 'e-Science' kind of approach are largely self-evident. However, in other communities, and particularly as you move away from the physical sciences towards arts and humanities end of the spectrum this is not the case. For many people the techniques of e-Science and e-Research are a 'dark art' for which training and support is essential.



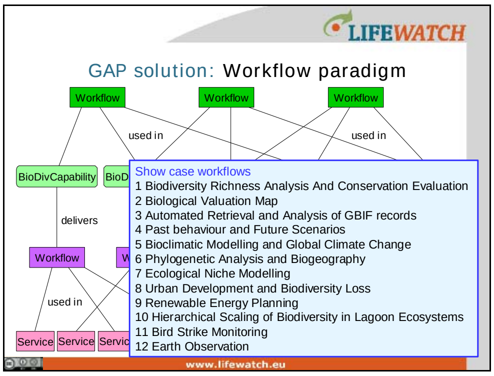
In biodiversity science the community is still at an early stage of development. Some scientists are pushing the boundaries of what is possible but many have yet to see the possibilities. Indeed, it requires a rather big-picture and visionary view to be able to conceptualise large-scale computationally intensive and data intensive research and to make that jump from what is routinely done today to something new that is hard to imagine.



Systems biology and the human physiome programme both have lessons to offer to biodiversity science when it comes to thinking about the environment we live in from a holistic systems perspective where linking and integration across scales – from genetic, through the scale of organisms, habitats, ecosystems and landscapes becomes the norm. As with the human body, it is not yet possible to model biodiversity at the global scale.

Slide 25

It is for this reason that we have adopted the idea of ‘capabilities’ to be provided by the infrastructure that can be composed in multiple ways to accomplish specific research tasks, such as these ‘show cases’. Such workflows, using tools like Taverna, Triana and Kepler, can replicate scientific analyses that are being used today but they also permit the flexibility to compose the capabilities into new workflows to answer questions and solve problems that have not yet been thought of.

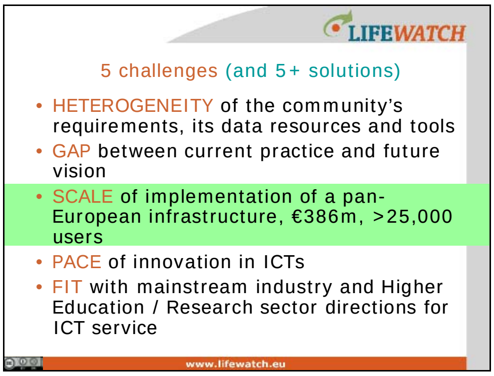


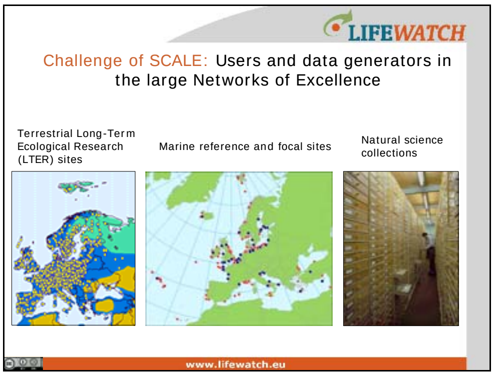
Slide 26,27

I spoke earlier about heterogeneity but now I want to address the scale of implementation as the next challenge to consider.

We don’t have an accurate estimate of the number of potential users but there are hundreds of research groups through Europe and thousands of individual scientists, not to mention policy makers, citizen scientists and students. In the early years we are aiming to support upwards of 25,000 users. That number might even be a considerable underestimate of the total in the long-term.

We are lucky that from the very beginning we have had a lot of good data providers: terrestrial networks, marine stations and natural science collections with all their specimens. All these together already form an important component. There are currently about 1800 terrestrial monitoring sites and 200 marine research sites across Europe. Hundreds of millions of specimens in natural history collections all over Europe are





gradually being digitised.

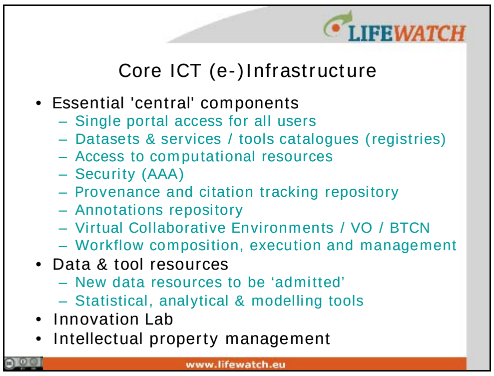
To deal with this challenge of scale, we use an

Slide 28, 29

approach that we call “thinking globally, acting locally”.

“Acting locally” in LifeWatch is a key feature that recognises the importance of the many small-scale communities of collaboration that exist in biodiversity research. LifeWatch engages with end users at the level of individuals, groups and organisations to bring about the emergence, added value and usage of the LifeWatch infrastructure. We anticipate and will encourage “islands” of infrastructure to emerge at the various levels of research groups, institutions, thematic networks, regions and nations. Over time, as individuals and groups cooperate and collaborate, such islands of infrastructure will organically grow and fuse with one another, much like railway systems have evolved over the years.

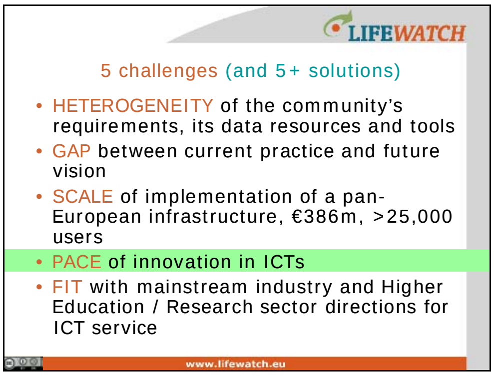
To achieve cohesion we *think globally* at the European level to set strategy, governance and guidelines that will bring about technical interoperability. We set the direction, we determine the priorities, we manage the LifeWatch product and performance, and we hold providers to account.
In many senses this approach mirrors the EGI / NGI approach you are adopting. I expect that National LifeWatch Networks will cooperate closely with NGIs in the future.



We will provide core ICT infrastructure for essential central functions and this will be implemented locally on behalf of LifeWatch. Some functionalities are listed here and you can see that we need access to computation resources of the European e-Infrastructures. In fact, our whole approach is based on the assumption that we build what we need – our service-oriented spatial data infrastructure – on top of common European e-Infrastructures. You will also see that there are a number of other essential components in that list that might be common to several ESFRI infrastructures and where some discussion around the detailed technical requirements will be needed with EGI once our construction phase gets under way.

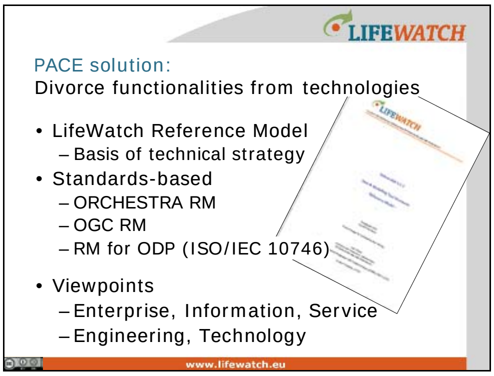
Slide 30, 31

Moving on, 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 new technologies, creating whole new tools and data platforms that depend upon them. Often this is done to solve an immediate need without thought for future-proofing and interoperability. The result is a jumble of tools and resources that cannot interwork, with each 'owner' acting as champion for their own approach as '*the best one*'. Faced with this situation and a long-term vision for interoperability, this presents a significant problem for LifeWatch. Technology moves so fast that it would not have been sensible to make technology decisions two years ago for a construction project that will not begin until next year. Had we made such decisions, would they have stood the test of time?

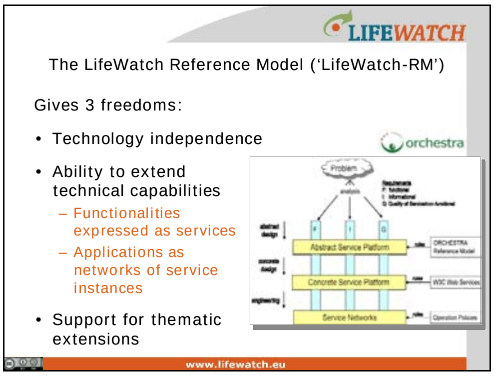


Slide 32, 33

We face the challenge of PACE by divorcing desired functionalities from their concrete implementation with technology. We have chosen to do this by adopting a technical reference model that will govern and direct everything we do. It is the basis of the LifeWatch technical strategy. It is standards based, being derived principally from ISO / IEC standards for Open Distributed Processing and Open Geospatial Consortium standards for the manipulation of information having a spatial component, which almost all biodiversity and environmental data has. The Orchestra Reference Model, upon which we have almost entirely based the LifeWatch Reference Model is an OGC Best Practice Document. Such documents are an official position of the Open Geospatial Consortium and thus represent an endorsement of the content by its 400+ members. ORCHESTRA offers a specification framework for the design of geospatial service-oriented architectures and service networks.

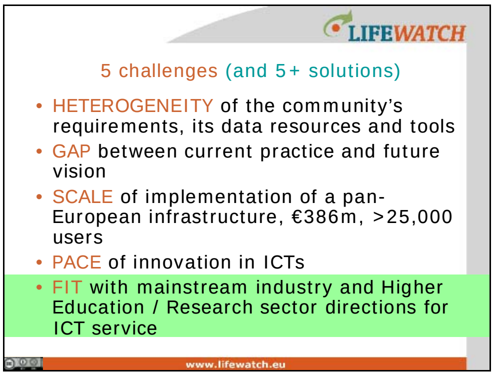


The LifeWatch Reference Model gives us 3 elements of freedom. Firstly, by separating the specification of the desired functionalities from the way that they are concretely implemented, we have freedom to adapt, to change the underlying technologies of the infrastructure as new products and solutions become available. Secondly, the reference model gives us the freedom to extend the capabilities of the infrastructure. By expressing resources and functionalities as services and treating LifeWatch applications as networks of service instances, we can achieve progressive implementation and re-use. Finally, the use of a reference model provides us with a set of implementation rules – meta-models and cook-books - for introducing thematic extensions to the

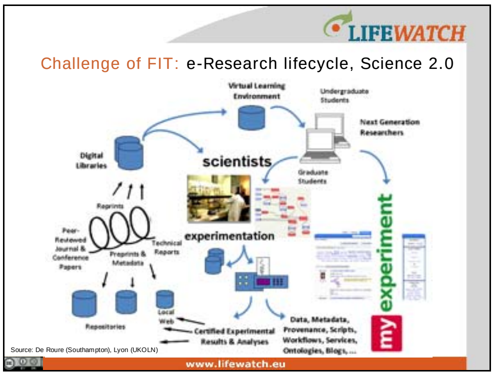


functionalities. For example, to support sub-communities of users with particular needs.

However, we must not forget that the LifeWatch
Slide 34 approach to e-Science, workflow, and collaboration should fit with mainstream developments taking place in industry and in the higher education sector more generally. Fit is the next challenge.

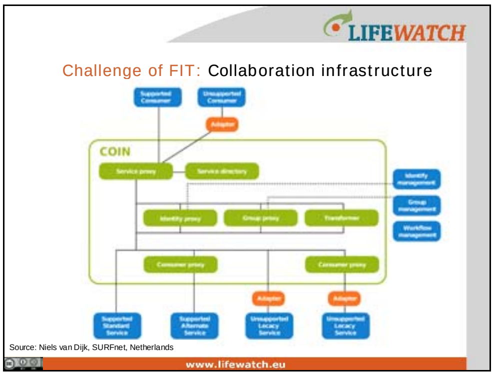


Computing services organisations in universities and
Slide 35 research institutes find it hard to support proprietary solutions. They are charged to push ahead to create modern working environments that contain all the elements of a complete electronic research, publication, teaching and learning lifecycle. Is Science 2.0 like Web 2.0? There are some people that argue that this is how it will be. Where do the current technologies of e-Science fit in this picture? What commercial products fulfil the need?

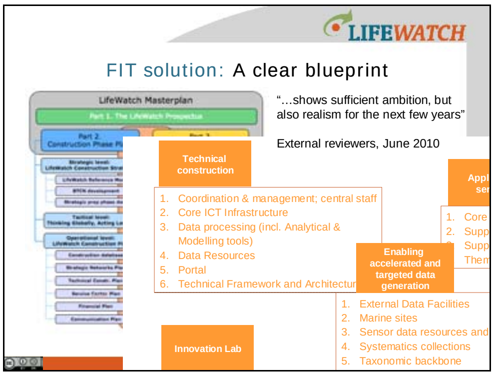


JISC in the UK and SURF in the Netherlands have invested heavily in experimenting with e-Research, virtual research environments and collaborative spaces. Recent publications from both organisations report on the technologies and lessons learnt. It is evident from those publications that there is potential; that commercial products are beginning to emerge but that there is some way to go to reach the necessary levels of functionality envisaged by e-Science and to make it all work seamlessly.

Slide 36 As we have already heard this morning SURFnet, with its next-generation COIN project, aims to create a collaboration platform for the whole of the higher education sector in the Netherlands. This will be based on open standards, federated identity, and a service spanning institutions. Users will be able to use internal, institutional services as well as third-party services. They will be able to use them in a collaborative fashion and to combine them to their liking. This is an interesting approach, yet it remains to be seen how its characteristics and standards-basis will match those of EGI and the ESFRI infrastructures, with which it should naturally integrate.

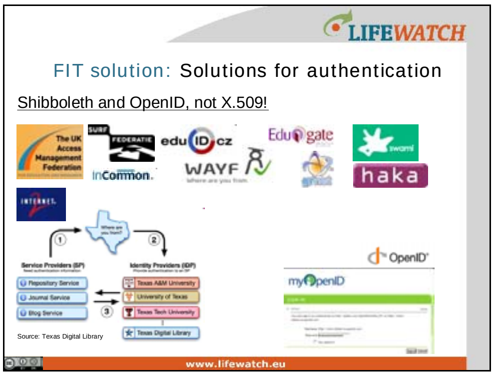


Slide 37 FIT is about having a clear blueprint for construction. We have that in LifeWatch and a recent review by independent reviewers concluded that ‘it shows sufficient ambition but also realism for the next few years’. We will bring our blueprint to life by working with the higher education and research sectors, with industry vendors, and with open source communities. Our approach will be to choose and adopt widely deployed generic solutions as far as is possible. We do not intend to re-invent the wheel.



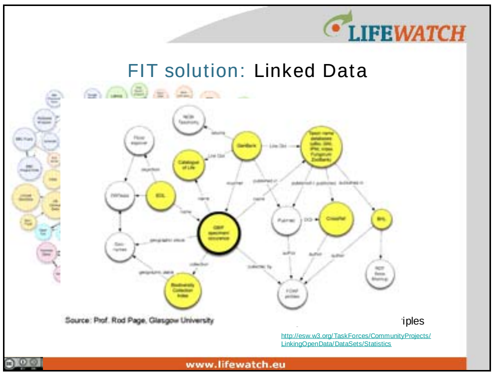
I give two topical examples: Solutions for authentication and for building relationships between data

Slide 38 It is evident that the higher education sector across Europe and in the USA is adopting Shibboleth as the basis of institutional mechanisms for authenticating access to electronic resources. OpenID is emerging as a solution outside of the sector – particularly good for those users without affiliation through their employer to a Shibboleth federation. LifeWatch will probably need both and will base its authentication and authorization



mechanisms on Shibboleth and OpenId, not on X.509 certificates.

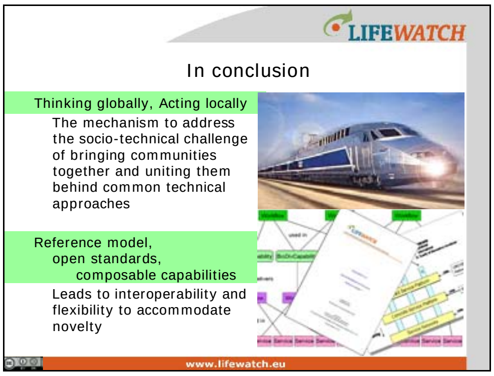
Slide 39 The Linked Open Data initiative has widespread support, including from governments both in USA and Europe, the BBC, the New York Times, the German National Library, and many more. At a recent count there were more than 19billion triples available and more data is being linked all the time.
Here's an example from Professor Rod Page at Glasgow University for how several major global biodiversity data resources could be linked, relative to one another and to other resources.



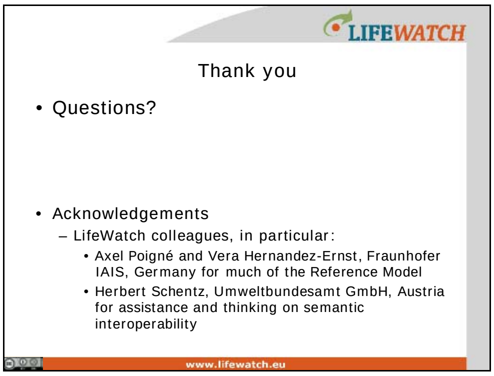
This brings me to the end of my explanation of the challenges facing LifeWatch. It's probably true that other ESFRI infrastructures face similar challenges and we are in fact working with environmental infrastructures on these aspects.

Slide 40

In conclusion I would like to think that, as with the high-speed rail network, thinking globally, acting locally is the mechanism to address the socio-technical challenge of bringing communities together and of uniting them behind common technical approaches to interoperability.
I believe that the Reference Model approach we have adopted in LifeWatch, its basis in open standards and the approach of composable capabilities will give us both interoperability and flexibility to accommodate future novelties.



Slide 41 Thank you for your attention.


LIFEWATCH
Thank you
• Questions?
• Acknowledgements
– LifeWatch colleagues, in particular:
• Axel Poigné and Vera Hernandez-Ernst, Fraunhofer IAIS, Germany for much of the Reference Model
• Herbert Schentz, Umweltbundesamt GmbH, Austria for assistance and thinking on semantic interoperability
www.lifewatch.eu