Notes for
http://wiki.pro-ibiosphere.eu/wiki/MS12_-_Workshop_on_mark-up_of_biodiversity_literature

For tweets please use **#pibber**
https://twitter.com/search?q=pibber&f=realtime
Please add notes at the bottom
Feel free to comment on existing notes, though

DAY 1

Ease of access to information ==> spent several days accessing information
how can we visualize what we actually know about biodiversity

there is no semantic markup yet; how can we turn this into semantic reasoning infrastructures

completeness
how can we handle large volumes

comments on Rod's piB presentation

use cases for markup:

- exchange & archiving
  - e.g. JATS (in PubMed Central)
- display
- tracking things (incl. forward linking)
  - citation
  - specimens
  - sequences
  - localities

protocols
authors
funding
eventually: ideas

Use cases can be addressed during the data enrichment hackathon that will take place in Leiden on the 17-21 of March:
http://wiki.pro-ibiosphere.eu/wiki/Data_enrichment_hackathon,_March_17-21_2014

Rod's questions:
*Do we need markup, or is indexing good enough?
*Markup for archiving, or better still, for version controlled editing (not for updates, but for fixing errors, including in markup)?
who's going to curate the edits/ pull requests?
*Does markup generate knowledge? If so, what have we learned to date from marked up literature?

What is the role of OCR (or better text capture) quality in the markup chain? How many errors can we deal with, and for getting answers for a specific question?

Where is the markup /knowledge graph stored? Use the markedup text or focus on the extracted content? Shall we aim at a central knowledge central graph

Annotations of errors: where to store them? PDF, markupdoc, DB?

Talk Dimitris Koureas, NHM

Slides at
http://figshare.com/articles/Biodiversity_literature_mark_up_Compelling_use_cases_for_Natural_History_Collections/928250

Use case 1

Assisted label transcription. Use literature markup to identify specimen records and match against the physical object

When you link labels to literature it gives you better info.

How can museums make markup part of their workflows?
How can markup facilitate label transcription?

759,000 GBP for three years for digitization. Is this a lot? What can you do with this sort of money? Is this a relevant piece of money within the NHM budget? Is this more than a commitment, a well-informed commitment?
We have to have an idea of what the costs are for markup, digitization and linking digitized material to the published record.

Probably an answer might be a policy that requests to only publications can be submitted that have all the elements in them digital and linked. This would allow to focus digitization and at the same time avoid creation of unlinked data.
Dimitris' mantra: "don't do the job twice" should be the guideline.

Different approaches to specimen label transcriptions
    manual, crowdsourcing, automated
    what if we could use literature markup (e.g. specimen catalogs) to link already transcribed specimen labels with the specimens
    that would also allow to find all studies that actually used a given specimen, which could help resolve some outstanding research questions related to that specimen

What's the impact of a natural history collection?
    value in itself
    value for/ in/ through reuse
    help in prioritizing, performing gap analyses
    metrics help with attracting funding (Nature commentary on how small collections can have a big impact)

How can collections of _physical objects_ be assigned unique (and permanent) identifiers?
    idea here is to allow tracking of usage already while the collection is not fully digitized yet
A solution to follow is also CASENt ant work which implemented http URI for their specimens, that also http://www.antweb.org/specimen/CASENT0104542 that also includes the image of the label as well as the transcript. how this might be used is for example in this treatement: http://plazi.cs.umb.edu/GgServer/html/8AD0DAEF2180649D27DBA7CE08E4FF93 and click on the linked material observation data, eg. CASC

GoldenGATE allows to make these annotations already now. But it needed to be made specific for this task.

Can we instigate change in editorial practice, such that specimen accession numbers have to be cited?
    See also:
http://www.pensoft.net/journals/zookeys/article/3178/no-specimen-left-behind-industrial-scale-digitization-of-natural-history-collections

We need to markup the literature and transcribe the labels. Only both will give you the full story.

Use case 2

Measuring NH collections impact
But this has already been addressed in the past by Rod Page on his blog. And this is also what pib is doing with the unique identifiers activities:

http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs

The semantic web works with URIs. There are 2 URIs. There are identifiers for the physical specimen

EU are interested in tools and products with societal impact

Talk Jeremy Miller, Naturalis: Scientific applications of markup

Estimate biodiversity of certain taxa based on occurrences reported in the literature
    could be automated

dashboarding:
    within treatments, provide an overview of the available data
        and make that updatable. example:
https://commons.wikimedia.org/wiki/File:Map_of_Paedophryne_localities_2.png

example scenario: finding a spider in the woods of Berlin in 1890 or 2014, when the location

commons machinery keeps track of provenance
    http://commonsmachinery.se/

Transcribing the label is not enough. Geolocation needs to be integrated with the whole markup.
If your users are modellers, they will need georeference.
If the uses are interested in identification, then that will not be the case.
It just depends on who your users are.

We need to have specialised projects, e.g. citation, catalogue alignment, platforms that allow users to improve the markup

Have a platform which allows to markup, offering the best tools for a particular task, eg georeferencing, ciation parsing.

Where should be markup and annotations be?

Data should be available when is needed, and in the form that is needed.

There needs a repository for collaboration (to structure the process), and a knowledge graph of interpretations.

Markup is interpretation.

You can fork interpretations, or separate them.

Different people have different interpretations and uses.

How does the incremental mark-up with the various repositories will work?
We won't achieve the vision without a cohesion of the instrumental process.
Can we have a convincing argument that we want to have a global infra?

Talk Bruno Jehle: Workflow and project management

What is what we want to achieve: Be clear about this to be efficient and able to deliver.

What is relevant.

People mix up things. They want to achieve a lot and they end in endless discussions.
Mechanisms for selecting materials

Step 1.
Content providers - Institutions
Selection and specification (list of docs, type of original)
prioritisation, how do we do this and who should do this?

Step 2.

- Originals. Digitisation as per minimal standard

- Quality control and OCR. Uniform data format, automatic OCR (if needed)

- Repository of base PDF (One needs to have clear what is in the containers (e.g. pdf))
    - Accessible on protected web server
    - Reference to single page
    - Filename convention

- Document structure
    - Page numbers, tagging of content

- Definition of elements
    - Text, tables, graphs and images

- Extraction of text
    - As per requested quality standard

- Final quality control
    - With reference to base PDF

Step 3 DOMAIN OF THE BIOLOGISTS! DON'T WASTE TIME WITH THE PREVIOUS STEPS.

- Database ingest
    - Scientific markup of content

Project phases
Proof of concept - Optimization - Volume production

Media standard

- Technical standards
- see ppt

Be efficient with the discussion and responsibilities.

OCR/text capture can scale up; it should be possible to scale mark up. We need to think of ways to achieve this, otherwise we will not be able to mobilise the legacy literature.

Mark-up can scale up. GNRD, as mentioned by William below, is very good at automatically detecting taxon names. For extracting locations from unstructured text there are many tools, with even better accuracy rates than GNRD achieves with taxa. For example, Yahoo!'s Placemaker or ThomsonReuter's OpenCalais as well as numerous packages you can integrate into your own code such as geodict for Python. If you want full mark-up to a schema like TaxonX or TaxPub, however, that is tricky. If you want individual elements then it is possible to scale mark-up. This all ties back to the fundamental question of ***user requirements****. What's the point of doing all this work? What mark-up is useful to users? What do we mean when we use the word 'mark-up'? Time to hand over to Rod to expand on this point :-) Myself, being a pretentious*

classicist I'd rather just ask: cui bono?

Maybe we cannot easily combine the approaches (programs) but we can make it easy for ourselves to combine the data.

Make data as interchangeable as possible through having all input/output **in a limited range of formats** only, say FlorML or TaxonX, so that future combination of tools is relatively easy because there is need to translate data among them.

Bruno Jehle competent in Phase 2, in Phase 3 possibly open doors to Indian communities

Is PDF the only input format, or can other formats be used, such as tiff, dejavu
double pages etc.?

William Ullate, BHL

Efforts and plans towards markup of BHL content

- no markup being done to date
- Mission has recently changed, now includes inspiration
- subject classification follows LOC subject headings
- Recently switched fom TaxonFinder to GNA (Neti-Neti)

There is article-level, chapter-level, treatment-level metadata at BHL, and some articles have started to be annotated by users.

Wikipedia BHL Art of Life

- example: https://commons.wikimedia.org/wiki/File:Diagram_of_Sparrow.jpg

Purposeful Gaming (based on Digitalkoot)
http://blog.microtask.com/2011/02/digitalkoot-crowdsourcing-finnish-cultural-heritage/

MiBio: Mining biodiversity

- http://www.nactem.ac.uk/DID-MIBIO/

Total size of BHL database is on the order of 70 TB. Takes too long to copy over Internet and have to ship around as HD.
This problem has been solved by CERN, who send around Petabytes per day to a network of partners. Don't remember details, but probably worth asking them (Daniel is in contact with them). GEOSS must have solutions as well with all the remote sensing data its members generate (eg NASA, NASDA, ESA)?!

Guido Sautter - The IMF Image File Format

How to map PDF and XML into one coherent framework, based on paragraphs and words and their positions on a page

A newly developed format by Guido (unfinished), allowing to store markup separately from the PDF, storing the bounding boxes together with the semantic annotations. *Have you seen djvu XML?*

There are annotation tools on PDF readers.
It is better to use the existing infrastructure, even if the end product is not that clean. Guido disagrees, PDF is too complicated, better to use self-developed format.

Is this sustainable?

No, it is not sustainable. PDF is complicated, but then it is also 'feature rich'. However, you probably only want to use a relatively small subset of those features. PDF is a widely accepted format that means there are many tools, and products, built on it. Furthermore, all mainstream programming languages have packages to allow your code to directly interact with PDF files. Why turn your back on a format and an ecosystem (for want of a better word!) that has tens of thousands of people working with it, supported by the resources of many commercial vendors, public institutions and funding bodies. I really don't see how you can match that scale of commitment.

All that said, PDF is complicated. If you really cannot commit to the learning curve required to realise the benefits of working with it, there are lightweight alternatives that have already gained some traction. The leading contender for a lightweight document interchange format is DjVu (http://djvu.org/) - alluded to above. I would argue again that joining an existing initiative begun in 1996 and still going, with an established toolset (see http://djvu.org/links/) represents a more sustainable future than starting from scratch and having to build, and then maintain, everything yourself.

Note, these comments also tie back to Bruno's earlier question about alternatives to PDF.

Plazi should provide APIs to analyze and visualize data in PLAZI. The value of adding data to PLazi should be more obvious

Plazi should provide tools that show for example the specimens collected by a collector.

Possible use cases for the hackathon?:
http://wiki.pro-ibiosphere.eu/wiki/Hackathon_%22Pimp_my_Data%22,_March_18-20_2014

visualizing Wikipedia edits: http://listen.hatnote.com/

Jeremy - create a tool that conservationists could use when analysing spider data.

Stand-off mark-up is far more flexible than in-line mark-up!.

See Roderic's post:
http://iphylo.blogspot.co.uk/2014/02/mark-up-of-biodiversity-literature.html

I sketched out the biodiversity "knowledge graph", then talked about how mark-up relates to this, finishing with a few questions. The question that seems to have gotten people a little agitated is the relative importance of markup versus, say, indexing. As Terry Catapano pointed out, in a sense this is really a continuum. If we index content (e.g., locate a string that is a taxonomic name) and flag that content in the text, then we are adding mark-up (if we don't, we are simply indexing, but even then we have mark-up at some level, e.g. "this term occurs some where on this page"). So my question is really what level of markup do we need to do useful work? Much of the discussion so far has centered around very detailed mark-up (e.g., the kind of thing *ZooKeys* does to each article). My concern has always been how scalable this is, given the size of the taxonomic literature (in which *ZooKeys* is barely a blip). It's the usual trade off, do we go for breadth (all content indexed, but little or no mark-up), or do we go for depth (extensive mark-up for a subset of articles)? Where you stand on that trade off will determine to what extent you want detailed mark up, versus whether indexing is "good enough".

Perhaps the confusion comes from a false conception that mark-up is synonymous with in-line XML tagging. In the wider NLP world, mark-up comes in many forms with stand-off being preferred over in-line primarily because you do not edit the source document. A second advantage of stand-off is that you can apply many layers of mark-up, selecting which one you want. An example for biology is the BioNLP format (http://2011.bionlp-st.org/home/file-formats) or better still from a usability perspective the brat derivative of it (http://brat.nlplab.org/standoff.html). Note, these stand-off formats also answer the question of how to record OCR corrections: the original text is marked-up as an 'Entity' and marked-up with a 'Note'. Hence, both original and corrected text are recorded in the stand-off file to be applied to the text in your viewer. Stand-off mark-up is far more flexible than in-line mark-up.

A further thought on the benefits of stand-off mark-up driven by Jeremy's mention of forking interpretations. If both source and annotation files are in a version controlled repository then it is possible to re-interpret the mark-up without affecting either the source text or the existing mark-up because you can fork a new branch of annotation files for your new interpretations. See the ViBRANT Corpus at https://git.scratchpads.eu/v for an example of this approach to mark-up hosting.

DAY 2

One responses from the questionnaire: “I'd like to know more about JATS and how we might be able to use it with legacy literature.”

So here we go with Chris Maloney (@Klortho), PMC
https://vimeo.com/86379933
http://jatspan.org/docs/BioMarkupWorkshop/#%281%29

Documentation framework jatsdoc
Descriptive vs prescriptive approaches
Human annotation of TaxPub DTD possible
NCBI RDF
PubChem just released their RDF

Bringing legacy literature into JATS
Mapping XML to RDF

--

Philippa
Important to talk about Money - revenue

Multiple efforts similar to piB are no longer around. Lack of sustainable business model is the likely cause.

How are the players similar?
They have defined products that users want
They design products to help users accomplish defined and verified tasks
They have concrete, scheduled deliverables
They have defined publishing processes
They think about marketing and competition
They think about costs and revenues (not necessarily profit)
They think about workflows
Most think about the world

Think who is going to do what, what is going to be the time framework, how much is it going to cost?

Royal Society of London, Knowledge, Networks and Nations

What are the use cases? what are the products that users want? Who are the stakeholders?

Making the argument that the BKMS is part of the infrastructure from the start of the partnership.

Distinction between use cases which could be enhanced through markup, and new use cases that only become possible (or doable) through markup

What already exists?

Biovel
==========
BioVel (EU funded project) uses biodiversity data to include it into workflows for e.g. ecological niche modelling, modelling changes over time by comparing old specimen data with new one, etc. The later can be achieved by using collection data and compare it with contemporary data

How and why to bring our data to BioVel?
We could support the BioVel workflow by opening up data from literature (floras, faunas, …) for inclusion into these workflows. This is of special interest as old data is usually underrepresented in such workflows, but urgently needed for certain tasks.

OpenUp
==========
The OpenUp (EU funded) project opens up digitized specimen media data to GBIF and Europeana. This is supported by a number of quality tools to ensure high quality of the related metadata. The projected ended in Feb. 2014 with ~ 1.5 Mio natural history objects mobilized to Europeana.

What could we do?
Media handled in OpenUp where mainly digitized material from collections (museums, institutions). We could add images from literature which is very rich in metadata which we already have marked up. We could link to their resources?

Or assure that content from iBiosphere can be found and linked to from Europeana. Eg markup dates so timelines could be made and data integrated across disciplines. Are there other such fields that could be used (names, locations, geographic data, diseases…)

Stakeholder: Europeana

EU-Bon:
==========

EU-Bon: providing observation and traits data
EU-BON task 3.4 has as goal to create a workflow to markup literature to provide input into the modeling activities. This includes observation and traits data.
Stakeholder: EU-BON

World Flora Online
==========
http://www.missouribotanicalgarden.org/plant-science/plant-science/world-flora-online.aspx

The UN Global Strategy for Plant Conversation 2011-2020 has as it's first of 16 targets an online flora of all known plants.
To reach this goal the World Flora Online initiative was founded. World Flora Online will require input from a diverse set of sources. Mobilizing literature data from existing floras and bring them into a format that is consumable will play an important role. This includes all data typically found in floras, from taxonomy over identification to traits data.

Stakeholder: World Flora Online, indirectly CBD

Provide trait and identification data
Stakeholders: Ecological niche modellers

Provide Collector data
Stakeholders: natural history museums

There should be a mechanism to connect the demands.

What is new?

How to bring citation to the collections?
Like libraries which do store huge amounts of literature natural history museums do store millions of primary biodiversity specimen. However, other than for literature the usage of

specimen for scientific research is not well documented and online databases for citations do not exist..
To know more about the usage of such specimen will help institutions in prioritisation and planning. Marking up and mobilising citations of specimen in literature will help to create such databases.
Stakeholders: natural history museums

Dates that are marked up will help historical research

Use case

scientific question, relevant literature, mark-up, CDM

Keep in mind the long-term goal of modeling the entire biosphere. What kind of data, tools etc. what are we lacking on the way there?

Why not cite the user requirements in our earlier reports and notes from the February Leiden meeting?

What can markup do?
1. Why XML vs indexing?
Question of granularity and what you can get out of a paper
2.
3. Rendition, extraction, indexing

Markup has to be driven by questions and demand.

Find prior descriptions

The power of aggregation

Difference between indexing and markup: The probability of making the right assumption is much higher through markup. This comes at a cost.

Markup of biodiversity information as a seed of markup for information relevant for other domains (consider tracking things like the history of insect trap use)

Link to Wikimedia and other resources?

Markup, semantically markup involves linking to external resources.

Does markup require publishing interfaces to its content?

Markup facilitates presentation of materials in multiple languages.

Markup allows for tailored notifications at fine-grained level.

Interest in traits rather than geographic data? Geographic data is more often incomplete in regards to geographic coordinates than specimen data - do we really now this? Are specimen data in collection more complete than the published record in regards of geographic coordinates.

Again, we need specific apps that allow markup of specific target elements (names, observation records). The markup should though come into a shared pool, eg use the Plazi serve

Does iDigBio do something similar to Dimitris' idea of using scholarly literature to transcribe labels?

There is a GBIF outlook document which may shape the biodiversity landscape over the next few years to decades ⇒ ask paper.

For hackathon ideas at Naturalis (incl. cooperation with Wikimedia), get in contact with Jeroen Snijders ??? …. with the organisers of the hackathon: Rutger and Soraya!