**Action Items**

All: Complete plus/delta feedback on meeting: <https://docs.google.com/document/d/1lxXgGK-5GR9G-tAf6_izj2p7H1y8EcURQXviwAI7aBo/edit?usp=sharing>

Connie: Policy level for annotations– Connie to work with Members and Affiliate on open annotation policy. Maybe even discussion topic at Annual Meeting (March 2018). Keri: Might be tied to SI policy on commenting on a blog

si.edu/terms

Katie, Mike, Ricc, Susan: Transcriptions \*First quarter 2018 to finalize approach for incorporating existing transcriptions into BHL. Discussions and recommendations should be actively worked on during the remainder of Katie’s residency (thru Dec 2017) to be able to incorporate her input. (Katie, Mike, Ricc, Susan)

Ahmed: Share the BA names data results so that Mike and Martin can review with Dmitry Mozzherin at TDWG (Oct 1-6)

Ahmed, Mike, and Joel – determine what we need for storage.

Martin and Ahmed – separate discussion on getting BA content into BHL

Figuring out what’s v.2 and what’s implemented in current BHL

**NOTES**

**Guiding Statement from BHL Chair**

* Sustainable
* Open Science
* Grow membership
* Connection to biodiversity repositories
* Connection to digital libraries

**BHL V2**

* Robust/nimble infrastructure
* Seamless software, tool integration
* Transcription & OCR
* Data extraction (names, species descriptions)
* LOD
* Easy accessibility
* Visual Resources / image search improvements
* Data security and preservation
* Mobile device improvement
* Image Delivery: IIIF? (solve problem of IA blocked; image manipulation)
* Incorporation of annotations
* Metadata cleanup / disambiguation (e.g., authors)
* Better exposure (ie Wikidata)
* Automation of tasks via machine learning

**Reality Check**

2018 Dues Budget:

* $275,000 avaialble funds anticipated
* $200,000 anticipated expenses
* $15,000 contingency (a)
* $48,000 unallocated (b)
* $63,000 potentially available at end of 2018

Reality is that we don’t have funds within Dues budget for bringing on additional full time staff to support new initiatives.

BHL Egypt has also offered to contribute to technical development.

We can consider grants and other opportunities in considering priorities implementation.

BHL Partners – so far we have been unable to tap into technical resources at those institutions

**Goals for this meeting:**

* Aspirational Goals for BHL v.2
* The kinds of things we want to do that we could seek funding for

Infrastructural funding vs project funding

In the past, grant-funded, project-based grants have distracted from core developments or larger infrastructural change. Project based grants will have to be tightly woven into infrastructure of BHL V.2. MRK and BHL EC share this perspective.

What is our approach? Do we define requirements, build over several years, then unveil?

Or put out a base implementations sooner and then incrementally add to it?

Preference might be to develop quickly, reveal and improve incrementally.

With this model, we would be freezing what we have and move forward as best we can.

**BHL Technical Vision Document**

<https://docs.google.com/document/d/14VixDfgl2-850ZmNhh_MSUfg3BKrX6W7hybJRMaNQXg/edit>

Harmonizing various aspects of system

How to make images available where IA blocked

Computable datasource – data exports

Where does BHL fit as a library?

Originally developed from a conglomeration of perspectives: Science + technology + library – some library things were left behind.

Goal was “to get stuff out there”

Now we can think about library services we might want to apply

BHL was originally just books.

Large segment of Members now want standalone visual resources

How do we revise data models and service models accommodate that?

What is needed for things like standalone Illustrations, maps

And should BHL do that?

If leadership chooses, Tech will follow

IA – still seen as an aggregation platform. BHL has a desire to deliver from elsewhere to enable service for places like China. IA remains relevant as a staging area, redundant storage space, failover storage.

Looking forward, having other storage opens up other options. For example, some things may not go on IA, i.e., if we have more restrictive content than we could have content that may not adhere to IA openness.

Managing infrastructure

SI servers are hosting content

Dark storage

Management of things

How to move forward?

Stop development on current platform, limit to bug fixes, maintenance.

Create requirements documents for BHL V.2

Get additional input from community

Make sure new req’ts support new standards like KBART

Adaptable platform, more modular

Best Practices in Digital Libraries

In the past, we had a Tech Director, for initial development. Chris Freeland. Then William Ulate came on and supported continued development. We have not had funding to support a Technical Director for a few years now.

TAG -- is the way we’re doing things efficient?

In the past, funding was at the grant level. With the current funding structures, EC is now more involved. Funders now are Partners and it is their voice that guides.

Exploring how we can have a more direct line of action while providing channels for input, along with clear chain of command.

**Core Platform Priorities**

Some are reflected in strategic plan and technical developer work plan

**Project Based Work**

Susan – In addition, grant to digitize Torrey at NYBG; in order to complete successfully we will need to get transcriptions into BHL.

MRK - Ricc Ferrante (SIA) is also eager to see transcriptions from SI TC in BHL

How we can more effectively manage technical development while keeping input and communication open.

NDSR Recommendations

Yesterday, residents reported on their work thus far on Transcriptions; Digital Library Best Practices; Enhancing Access to Illustrations; Collections Analsys; User Needs

**BHL Egypt**

* BHL Synchronizer
* Online Archive OnA
* Digital Assets Factory (DAF)
* Digital Assets Repository (DAR)
* Hadoop layer
* Website

*Workflow*

IA hosted at Bib Alex

Synchronizer pulls from IA to online archive

Then DAF download extract publishing backup

Stored in DAR, includes book viewer and other nodes

DAR: Fedora & 4 Store

DAR: Access Layer –

HADOOP – metadata enrichment, sci name extraction, solr indexing

Connects to Backend website

Synchronizer – developed at Bib Alex, Synchronizer downloads from IA and stores on OnA (the BA archiving system)

OnA – Online Archive. This is a low cost scalable system. Running specialized BA software. AIP ingested is mirrored at least once to provided redundancy

Archive (store) BHL books

DAF - BA dev’d, Open source. Manage workflows, OCRing and so on

DAR works with METS and MODS, DAF is only way to ingest to DAR

Download 🡪 Extract images from jp2 and jpg, to gen PDF and DJV 🡪 Publish 🡪 Backup

DAR is the repository:Books, images, manuscripts, serials

Provides free public access

BHL is second largest collection in DAR
DAF ingest books to here

Stored in different copies

* Original;
* zipped;
* processed images

Fedora + 4store

Hadoop index book metadata and indexing

DAR - get new and download metadata fedora

Hadoop

Doing all the extensive processing in a very efficient and scalable way – metadata enrichment, generating scientific names, indexing

Metadata enrichment – geocoding locations, Lat and long from addresses

Designed to use multiple APIs at same time

Respect max limit for each API

Languages – full names from lang code, ISO

Communicating w EOL

Three Solr cores

* Book metadata core
* Scientific name cores
* (EOL, book ids)?

BA Taxon finder

Intelligent algorithm to search BHL for all taxon concepts

Link founded taxon concept to EOL

Website Features

Search features – title, lang, author, location, tagged species; auto-complete in search box; sort result by title, number view etc; faceting; save search query; create collections by user; Rating; User reviews; book metadata displayed; related books; list of other books viewed by othr users;

Geogrpahic – map visualization

Scientific name – uses EOL to show hierarchy tree;

Highlights

Implement workflow

Publish 57,000+ books on bhl.bibalex.org

Enhance website design and user experience

Bib Alex uses the Scrum process

**Discussion**

**Names -** Taxonomic name finding – BA developed own instead of Global Names? Yes

MRK – we are committed to Global Names algorithm tool

Mike – it’s faster. It’s reliable. Right now serving us well.

Comparison of names found – Ahmed found a significant increase of names available

Mike: Don’t know if GN getting all or right names

Connie – we’d have to somehow vet what the two tools are doing

Dmitry Mozz. Will be at TDWG (Oct)

\* Action Item: If we can have data by then we could share with him to discuss

William: Not using names BHL found; Using your own algorithm,

Ahmed: Yes. Using only EOL.

William: GA uses EOL along with about 20 other lists

Ahmed: Gets updated version of EOL about every 3 months, and updates.

Susan: Looking now and it appears serial publications are not grouped the way they are in BHL.

Let’s step back – are we developing on top of what Bib Alex has or we asking to start new? Or building on top of what BHL has?

**Storage**

What do we need for storage?

What about the proposed options MRK sent around?

No comments.

Page images – we store in Isilon

Joel and Ricc – not fast and having capacity issues

Storing images – can make showing things faster, but you’re really showing different images. Like with maps. More storage required

MRK: SI joined IIIF as full member. SI is doubling down on IIIF.

Susan: What about amazon web services? High capacity and easily scalable.

With SI, it took a lot of time.

Joel: Amazon – probably very expensive 150 – 200 TB of data. BHL is already at 110 or 120 TB

Keri: it’s cost prohibitive. Cannot use AWS.

Don’t think SI has not settled on an image store and IIIF. Think we’re still exploring and testing.

MRK: Currently not paying for SI servers. They donate platform. Sometimes you pay initial and they replace. For BHL, everything came free. Cost was time. We’re behind firewall. Since we don’t know overall what image platform will fall.

Keri: SI will be moving to new repository of data tens to hundreds of petabytes of data; 300 TB of BHL is going to look tiny.

5 years. It has to happen

Ahmed, Joel, Mike and maybe William.

What we will need as a place to put stuff.

Very liberal estimate of storage needed.

If we open to more types of things.

Maybe even a petabyte….

2007 BA integrated into San Fran IA. Separate discussion on getting that content into BHL

At one point, there was a limit of how much you’d put in. Copyright module and © dept, books that are no longer in © This is not where the difference in number comes from. Difference in number, Comes from pipeline of process

**Transcriptions**

1. Getting existing into BHL
2. BHL method for collecting new

1. Existing

How do we want to transcriptions FOR MANUSCRIPT MATERIAL to live in BHL?

Structure – need a structure for putting in to normalize. As far as SI transcription, relation established to item probably at ingest

NYBG grant – laying transcribed txt exactly where OCR is now. Delivered, it’s a low budget way.

For example, consider a transcribed version – 100 pages in field notebook.

Each file would be a transcription for each page. And a zipped file for whole item. Sequences need to match

After, the unzipped in place in BHL

Would like to see scientific name extraction run on transcriptions

SI collects transcriptions as single xml file, easily parsed. Can be mirrored to page image. Multi-page PDF, we lost page-to-page comparison.

NYBG – one-to-one transcriptions per line.

From tech view, how stored is important. SI Transcription, handles as a separate line in xml file.

Can export plain text – might be easiest but we can also explore xml

Both ASCII

How do we normalize?

SI is using tags inside for certain things.

Revisit those recommendations in a timely manner. Katie can be a resource for that review

Versioning; version control

If someone transcribes a book, better OCR, another OCR or transcription. Are you keeping the old one, does it get overwritten?

Structural and Policy things; So this might be a policy thing

Annotations for identifying errors in transcriptions; users find it frustrating to see errors in transcription and not be able to correct.

Katie can clarify issues, solutions, and then open to wider group to formalize as workflow

How to transform into a workflow with stuff coming from multiple sources?

Connie: How do we prevent overwriting from IA?

Mike: Will not be overwritten unless someone adds a page to that book. Then it gets overwritten.

Thought we had found a way forward so maybe let’s not start from scratch

\*First quarter 2018 to finalize (Katie, mike, ricc, susan)

Part 2: NewDo we want a BHL transcription Module?
Katie has recommended one

Maybe building a version of a FromThePage in BHL context for people who may not have access to run their own transcription projects

If we want plain text as transcription files--If volunteer wants to mark up geographic scientific names, do we need to think about a place to store that?

First: Taking in as is now

Two: If we design our own, think about how to take advantage of incorporating that

Keep it. Even if won’t have the ability until later. Have it for later.

Plain version will be good for name finder

Structure is very valuable

Part of discussion should be vocabularies.

\*Answer at end of meeting

Timeline and priorities for building BHL transcription module

To build this and adapt module, what is that as a priority?

We can already bring in transcriptions

What is community management?

Could be link from page in BHL to site where you transcribe

Built in community engagement tools; so many opportunities you don’t want to encumber in portal

**Annotations**

1. Historical: Darwin library annotations -
2. Corrections: This species is now that species
3. Notes: I find this page interesting
4. Linking
5. Tagging – just a term

Marginalia versus user annotations can be managed different.

Might be same thing in terms of how stored.

Just needs to be searchable as different kind of tag

Who gets to see them?

Hyptohes.is and AnnoTate – the trolling aspect of annotations. Negative things around annotations.

Public? Private?

Policy issues, terms of use

You need to manage that community

Hypothes.is has a FoaF model

Annotations on BHL Egypt – Discourse and Disqus, open source

Like social media

Both of those have good access rights system

Disqus was on BHL but ran into terms of service problems

Connie: IIIF as an annotation platform?

MRK: This would be curated annotation by staff

Annotation feature of IIIF at SI? Others are interested but mostly in curated way. Annotations can be used in multiple ways. Also developing to work with authenticating. Logging to make or view certain annotations

\*Action Item:

Policy level for annotations– Connie to work with Members and Affiliate on open annotation policy. Maybe even discussion topic at Annual Meeting. Keri: Might be tied to SI policy on commenting on a blog. See also si.edu/terms

Priority level for implementing annotations?

Search

1. Usability – some search features we already have users don’t know about them or how to use them
2. Metadata searching
	1. Including faceted
3. Full text searching
4. Browse

Susan: Search examples next to search box in IA; FAQs or Help right by where it’s needed.

Joel: if have to explain it, might not be user friendly enough

Standards for data, For example. Years

Cleaning up data might help with search accuracy

Users would like to Sort by date added. They search for same thing over and over and just want to see what the new ones are.

Are they searching for something specific or broad scope of things?

MRK: I think they’re mostly specific searches.

CAS: There’s a mix of user groups. Individual website visitors vs API users.

**Demo of full text - Mike**

More results!
Faceting!
More to come!
--Applause---

**High-level things**

Discuss in terms of Technical difficulty/ tech specs/ data model changes

Recommendations

How might fit into things BHL Egypt’s model already or could accommodate

* Usability studies
* Improve search and browse
* IIIF
* Transcriptions & OCR Corrections
* Data exports and documentation
* Annotations
* Incorporate filtering scientific names, etc.
* Names (author, scientific, geographic, common)
* Enable contributions to Wikidata
* Digital Preservation

Matrix of simplicity (Difficulty on one axis, Payoff on the other)

What are your expectations? How do these fit in?

Susan: Transcriptions- adding existing are Urgent. OCR not as important because we have something. Might be 90% accurate

Filtering – full text will help a lot with addressing search feedback

Also still need metadata cleanup, years for example. There’s a lot of variance. Sometimes it’s in the volume field and the year field is empty. Finding an automated way would be ideal. Looking for patterns.

There are processes now to extract from volume field. However, a good third still don’t have year metadata, 67,000 at item record level.

Added extra fields to hold start and end dates parsed out of volume string. For KBART purposes

Additional dates could be populated.

Archival stuff will have dates at page level.

Worth a look.

Some things in place to help with data cleanup for Years. Needed for other things, too. For example, Author names.

For transcriptions, are we talking BHL 1 or priorities for BHL 2?

BHL 2 is adding in a tool; BHL 1 is adding existing transcriptions.

Investment in underlying data. If we invest in that, it carries from 1 to 2.

If we’re going to freeze how does that fit in?

Requirements are defined. Just a matter of finding time to do it. We may now need to add to requirements. But we can move towards a proof of concept.

No code freezing until 2018.

When would we be looking at for that?

First we need to develop what core platform will be.

Components for planning/consideration

* Basic Core Infrasturucture
* Things that require people
* Features

Infrastructure Needs

Then Features

Then can determine the Platform

What is likely to require data model change?

What if we want to move towards Big Data? That might mean things need to be different.

If we want future flexibility

Flexibility – chicken and egg

Things to plug in at a later

External APIs

Requirements

Define what is needed in terms of sys specs

Then Features

Then For user

Then we decide on infrastructure

Hadoop

Bib Alex Moving away from MySQL

BHL is now SQL server. Need to move into Open Source

Maybe Hadooop for business, can be used for digital preservation

iRODS – manage data itself

Need to be very specific about requirement

Will need to revamp technology

Ahmed: Scrum = best approach for this

First, need to decide the roles

Product owner: Features organized by requirements; requirements from business to tech team;

Scrum master: person manages scrum process, organizes checkins, etc;
Team – software engineer, developer etc.

3 meetings – 5 minute standup meetings checkin – today and yesterday. Sprint – planning meeting, tasks for 2 weeks

Then what has been done what has not been done

Artifact – product backlog, they’re using JIRA

Need to go into each one before deciding storage

iRODS perfect for open data management, but can’t decide until we have the requirements

At what level do we want to contribute to WikiData?

Annotations – need to write what it is actually supposed to do for each type

Might write need autocomplete feature, need to write specifications.

Ahmed’s Team:

3 software engineers. Full time dedicated. Can assign more based on scope

Also have dedicated designer team

Let’s start from scope before assigning a team

What else might be missing?

Richard: Article level metadata?

Mike: Requirements are defined are for that.

BHL Egypt Work Plan

Bib Alex team needs access to examine current application

Can offer Data validation

Ahmed: We can agree on then top level. Then move towards refining details and work.

Note: 2b and 2c are the same; should be merged

Martin will be making edits to the Work Plan document

MRK: Need to merge development stack, wherever it is.

It’s one system

It would be great to incorporate technical abilities of Bib Alex team into design and implementation. Are there things that BHL Egypt can work on now while we figure things out?

What are some things we should be looking at around data?

WikiData efforts are very preliminary, ad hoc at this point.

Europeana – large scale data donation opportunities we might want to look into

Computational data store of BHL

Developing more programming packages to better expose and document APIs

Susan – like digital humanities. Word frequency, etc. But digital humanists might not be primary stakeholder

Didn’t hear users requesting these features

Going back to Mining Biodiversity, though, there are related needs. For example, can you get to wing length = 4 cm

Parallel to DH concepts, traits.

Maybe identifying habitats or temperate range, adding other types of structure

Connections

What makes it valuable to GBIF and users like that. Species occurrence data, surveys, checklists. How to pull out species and geographic data, in DwC format??

KM: If you’re going to crowdsource transcriptions – human could put common name w scientific, could be used with machine learning.

Do we need to add a work package for data things?

System user survey will provide direct input. Are we speculating?

Connie: The surveys are supportive. But we don’t need to wait. We know already that text mining is valuable.

Are we starting from scratch? Or building off existing?

We’ve had developers. Haven’t had designers. That was BHL Australia They did full UX UI. That would be really useful to have access to.

MRK: BHL is book oriented. Are people going in for book experience? Should there be an alternative view for people who want to go straight to a species description?

Users looking for relevant literature, find original description.

Problem with that is that species descriptions aren’t just on one page; can skip around.

Ari: Have been conducting a lot research around the main tasks that researchers do, maybe could pass that along to BHL Egypt designers.

If we’re talking about re-vamping the book viewer, do we want to take a broader view? Change book viewer to digital object viewer?
yes

Ari’s recommendations might go beyond just book viewer; other aspects of navigating the site

What is Backstage / how does it work?

Backstage uses NAACO

Taking creator records and making available to Backstage, they would then cluster then. If look like same person. Then same human intervention to review.

BHL catalogers want to add VIAF IDs when possible. In data model now. They’ve started to do that.

What is the hold up?

There’s enough work for a grant proposal. Could be 2 half time people for a year. If throw in WikiData.

Fee from Backstage. Then work for Mike for new creators, matching needs to be enhanced

OpenRefine

Joel had intern looking it at. Decent results. Could be applied to BHL

Katie – tried it using the manual. Metadata isn’t in a place for it to work very well because of years. It does require a lot of manual work for those without that.

Known problem around date of birth and death, esp with things like circa

Do we want See Also references in BHL?

Linneaus to Linne, for example.

From a library perspective, yes.

But the system has the ability to group them together for the user and presented in an intutitive way, we no longer have the same need for See Also. Current data model supports infinite names, spellings. Could have one Record with multiple spellings. Alternative spelling are our friends if we have a way to store them. Not labeled as See Also.

**Product Owner**

It’s so valuable for developers and engineers to have access to detailed and immediate knowledge from a Product Developer. At EOL, this is Katja and Jen.

Scrum master – meetings on time, sprint planning, organizes 3 5-minute standup meetings; working with Jira; essentially project manager.

Would the EC be willing to have Product Owner manage the decision making of technical priorities of implementation?

Key position – Product Owner

Given Responsibility to implement changes and ordering tasks. If bug comes in, they make those decision of how fits into priorities.

Martin would the equivalent of Product Owner

Senior Director – says to get it done. Not day-to-day management of it.

Product Owner is responsible for getting it done.

Buy-in from Members is important, but need to determine at what level?

From there it gets carried out.

Connie: Interested in making it efficient, improving process. Still not sure how a Product Owner fits in. EC and Members are the owners.

Doug: Agree that need Members’ input. Just at a high level. Just having a more clear goal of end vision. Within that, no need for it in the day-to-day.

Define decisions that can be made without consultation, just informing

Being beholden to a committee that meets every 3 weeks definitely does slow down momentum. Maybe issues don’t need to be addressed just during meetings. Maybe an email would suffice.

What kind of change really requires discussion vs notification or just email check-in? BHL EC can generally be more quick via email

Susan: BHL tech team could have better communicate with Cataloging Meetings and Collections Committee. It’s incredibly important; they’ve been grateful someone from Tech group participates in their calls.

Because we are a consortium model, we don’t have control over the resources participating in BHL the same way an individual institution would.

We could try to ask things of Partner; but we have to know they can say no and be prepared for that.

If we had a clear roadmap, if we had a product owner that defined minutiae, then there’s no need to consult after roadmap approved.

Focusing on users – the why, the problem we’re solving. Can help build buy-in.

Need to consider when major changes to user needs come up, how we would handle that.

Connie

We need to focus on Better communication. There needs to be more careful review of Tech and various committees by EC.

Think of it as a roadmap, we don’t need to decide on every turn. Just that roadmap makes sense. If overall roadmap changes, EC needs to know.

Ricc

Product Owner may realize more intensive and frequent input.

If there are new recommendations EC can veto, otherwise move ahead

Connie - We should empower MRK to know when something really needs to come to EC.

If a road map like Tech Priorities MRK shares at Members Meeting is approved, nothing else may need to be brought to EC the rest of the year in terms of Tech priorities. Unless something fails or something new comes up that falls outside of those Priorities.

So from high level goals which are mostly mapped out. Going forward we determine how will be rolled out with BHL Egypt across work packages, which work to stay with SIL and MBG.