

Annotation Workflow In Natural History Collections

Report for the SYNTHESYS II project

Network Activity 3, Deliverable 3.7

Jörg Holetschek
Botanic Museum & Botanical Garden Berlin-Dahlem
Königin-Luise-Str. 6-8
14195 Berlin-Dahlem

Table of Contents

Introduction.....	2
Traditional Annotations in the Herbarium Berolinense	4
Virtual Annotations in the BioCAsE Portal Family.....	7
JSTOR Plant Sciences	12
The JSTOR Annotation Mechanism	12
The JSTOR Annotation import mechanism	16
Filtered Push.....	16
Acknowledgement.....	17

Introduction

Natural history collections worldwide house huge amounts of preserved organisms, which have been gathered by collectors in expeditions throughout the world over the past 300 years. The collections continue to grow by constant addition of vouchers gathered by researchers and naturalists active in systematics, ecology, biogeography and applied fields such as forestry, crop genetics and pharmaceutical sciences. Collection-based research contributes new information, which in turn is added to the collection objects and their documentation. The collections thus represent the result of innumerable working years of professionals and volunteers gathering specimens in different regions of the world, documenting their findings, and annotating existing preserved specimens.

In the traditional work with specimen collections, annotations play an important role. Herbarium sheets, for example, usually have labels attached with typical information on the specimen: the collection date and locality, the collector's name, the name of the identifier, the date of the identification and of course the identification result itself, a scientific name typically on species or subspecies level, sometimes on a higher level. By nature, some of the information (for example on identification and gathering locality) are error-prone, so it is not exceptional that scientists disagree about these items for a specific specimen.

If a scientist working with a given herbarium sheet finds out that one of the information pieces printed or written on the sheet label are incorrect in his opinion, he will attach an annotation slip expressing his doubts and stating the supposedly correct information. Under no circumstances he will change the original label, since he can only express his opinion. It is up to the view of scientists working with the sheet in the future to decide for themselves which information they deem correct. The curator can later review the annotation and decide how he would like to react upon this annotation. If he considers it irrelevant or simply incorrect, he can just ignore it. In case he agrees to the judgment of the annotator, he will add a revision slip to the herbarium sheet. This might also lead to an update of the herbarium catalogue (or herbarium database, if the catalogue is stored digitally) and moving the sheet to another folder or another shelf.

The figure below shows a typical specimen sheet with annotations. The specimen label is in the far left corner at the bottom; the slips above are annotations. The topmost slip on the left side

designates a revision, so the curator has accepted an annotation and changed the original identification.



Within the frame of the BioCASE project¹ (Biological Collection Access Service for Europe) and the SYNTHESYS I and II projects² (Synthesis of Systematic Resources), substantial basic research on annotations was done in the past. This involved

- The development of the SYNTHESYS Annotation System Prototype,
- Its integration into the BioCASE data portal family,
- The specification of an improved annotations storage mechanism, and
- The development of a mechanism to import annotations from the JSTOR platform³ into the Herbarium Berolinense⁴ (herbarium at the Botanical Garden & Botanic Museum Berlin).

Based on this work and in close cooperation with the SYNTHESYS team, a project for “a generic annotation system for primary biodiversity data” funded by the German Research Foundation (DFG) was initiated at the BGBM. This report documents mainly some of the findings of Okka Tschöpe and Lutz Suhrbier for this project (still running). **Important Note: Since some of these findings will be part of a publication still pending, this report should not be made public before six months after its delivery.**

¹ <http://www.biocase.org>

² http://www.synthesys.info/II_na_3.htm

³ <http://plants.jstor.org>

⁴ <http://www.bgbm.org/BGBM/research/colls/herb/default.htm>

The goal of this report is to document the work with annotations in today's natural history collections. This involves both the traditional, physical specimen-based handling and the management of annotations in data portals such as BioCAsE (Biological Collection Access Service), so-called virtual annotations. For this, the workflow for annotations has been analyzed in

- The Herbarium Berolinense,
- The BioCAsE data portal⁵, which uses the SYNTHESYS Annotation System Prototype,
- The JSTOR research platform, and
- The Filtered Push network⁶.

Traditional Annotations in the Herbarium Berolinense

This basic annotation workflow reflects the classical "offline" way of annotating specimens available at the specimen archives in the BGBM herbarium. In general, specimens are mounted on paper herbarium sheets and archived shielded by a boarded folder. The specimens are sorted taxonomically.

Any original description is either written by hand directly on the sheets or printed or written on labels affixed onto the paper sheets. All digitally recorded specimens can also be identified by barcode labels affixed to the sheets.

Annotations may be written by hand directly onto the sheets, written or printed on note sheets which may be affixed on the sheets or loosely appended to the boarded folder. Sometimes, annotations may also be written and appended on photocopies made of the original specimen sheet. Commonly, an annotation includes the name of the herbarium, species, subspecies or variety name, including full authorship, the name of the person who made the correct name determination ("Det."), date of determination, determination source and the investigator's name and institution. If plant material is removed, the type or purpose of the study and the reason for removing material should be given.

A scientist who wants to study a certain taxon will first need to request access to the herbarium. Access is permitted by the collection curator if the scientist has a reputation or is associated with relevant institutions or authorities. Equipped with such a permit, the scientist needs the information on where to find the specific specimen. He then can enter the herbarium and compile specimens in consideration of the individual research interest. Specimens can also be sent to the scientist. While analyzing these specimens, a scientist may decide to append annotations to the specimen sheet as described above.

Usually, the following kinds of taxonomy-related annotations can be found in classical Herbarium environments:

- New determination: if a specimen is undetermined or determination is wrong,
- Verification of identity: if determination is confirmed,
- Update in nomenclature: adaptation to new taxonomy, original name is now a synonym of the new name,

⁵ <http://search.biocase.org/europe>

⁶ <http://etaxonomy.org/mw/FilteredPush>

- Classification: if taxon is assigned to a new group,
- Type designation: type annotations designate the kind of type (holotype, isotype etc), the basionym and the complete literature citation. If the current name differs from the type name, the specimen should also be annotated with the currently accepted name.

After having finished the examination, the scientist returns the specimens back to their original place within the herbarium archive.

In general, appending annotations to the specimen sheets is the only form of communicating research results or error corrections to the taxonomic community. Thus, neither the curator nor the scientific staff of the Herbarium will get immediately informed about the appended information. Consequently, annotations represent information holders and transmitters enabling the scientific discourse about specimens. Typically, this cycle will only be suspended if researchers decide to scientifically publish their research results along with a documentation of their sources (e.g. annotated specimens). That way, the herbarium staff may take notice of scientific advancements regarding specimens stored in their archives.

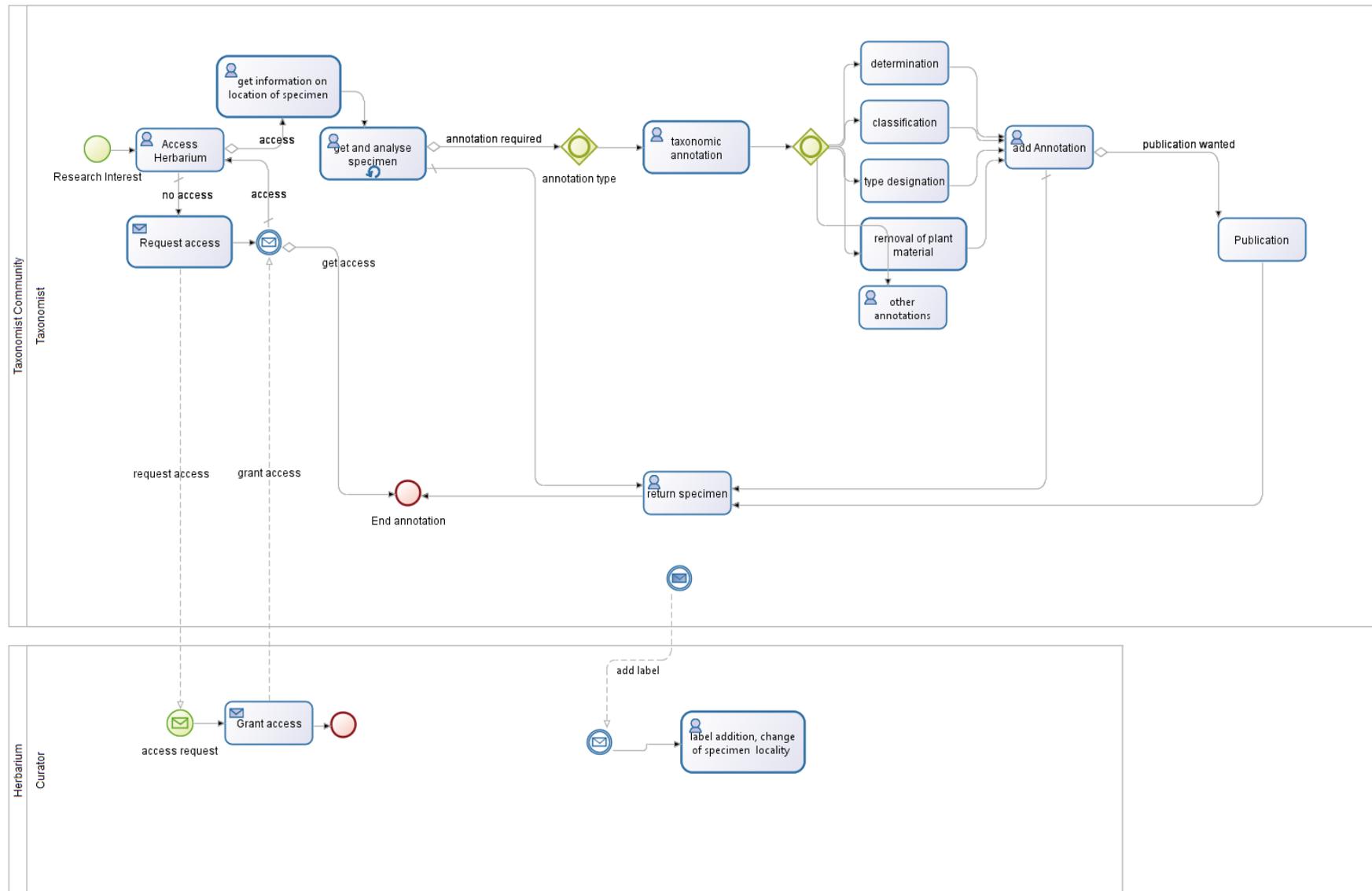


Figure 1: Annotation Workflow in the Herbarium Berolinense

Virtual Annotations in the BioCASE Portal Family

This annotation workflow reflects the current way of annotating specimens online through the BioCASE portal. This allows users to search, select, examine and analyze specimens online using a common web browser. For performance reasons, a search index will be cached and updated regularly at the portal server.

Therefore, the portal server requests the registered data providers for the required basic information on a regularly basis. That way, it is not necessary to request a complete specimen dataset from a data provider before a user explicitly wants to examine it. In both cases, the portal server communicates with data providers using the BioCASE protocol.

The BioCASE protocol enables data providers to join the network by mapping their specific individual database schemas to the XML based data exchange formats ABCD or Darwin Core. Based on these data formats, the portal server's web interface prepares a presentable data representation allowing users for detailed specimen analysis.

Registered users may enter annotations as free text comments at the bottom of the web page representing the annotated specimen data. Further on, annotations may be done by modifying the XML data (ABCD or Darwin Core) generated by the data provider. After having finished the annotation process, the web interface transmits any annotated data to a dedicated annotation server. That server stores the annotated data and comments together with the current version of XML data received from the data provider. Also, it notifies the related collection manager about the latest addition by email. While the annotation's free text comment will be managed by the annotation system, a second text field offers users to send an unrecorded, private email to the collection manager separately.

When an annotated specimen is viewed on the BioCASE portal, the user will be presented with any annotations. Thereby, any modifications in comparison to the original dataset will be visually highlighted through application of different (text) colors. In this case, the original dataset means, that the annotation is compared with the revision of the original dataset at the time the annotation was created. That is, the current content of the original dataset may have changed in the meantime.

After having received email notifications about the recent comments or modifications to data sets, the relating collection manager hopefully starts reviewing the affected dataset in a timely manner. Thereby, it is on the collection manager to decide on accepting or rejecting a proposed modification. Accepting it means that the corresponding dataset has to be updated within the collection's database.

Though, the next time the updated specimen dataset is requested via the BioCASE protocol from the collection's data provider, this new revision will be transmitted. Due to update cycles of BioCASE portal's cache index, there might be a gap between the information cached in the index and the information delivered by a direct request to the collection's data provider.

Currently, there is no way for the collection manager to create replies to annotations which will be recorded by the annotation server other than creating a new comment on the annotation. Nevertheless, since the notification email contains the email address of the annotating person, the

collection manager may send a private email to him. Also, there is no possibility for collection managers to notify the annotation server or a subscribed community about new revisions in his collection.

Common annotations that are made include

- Correction of typos,
- Correction of coordinates (georeference),
- Correction or specification of localities (Country),
- Addition of author names for taxa, and
- Specification of basis of record (observation, specimen, living collection).

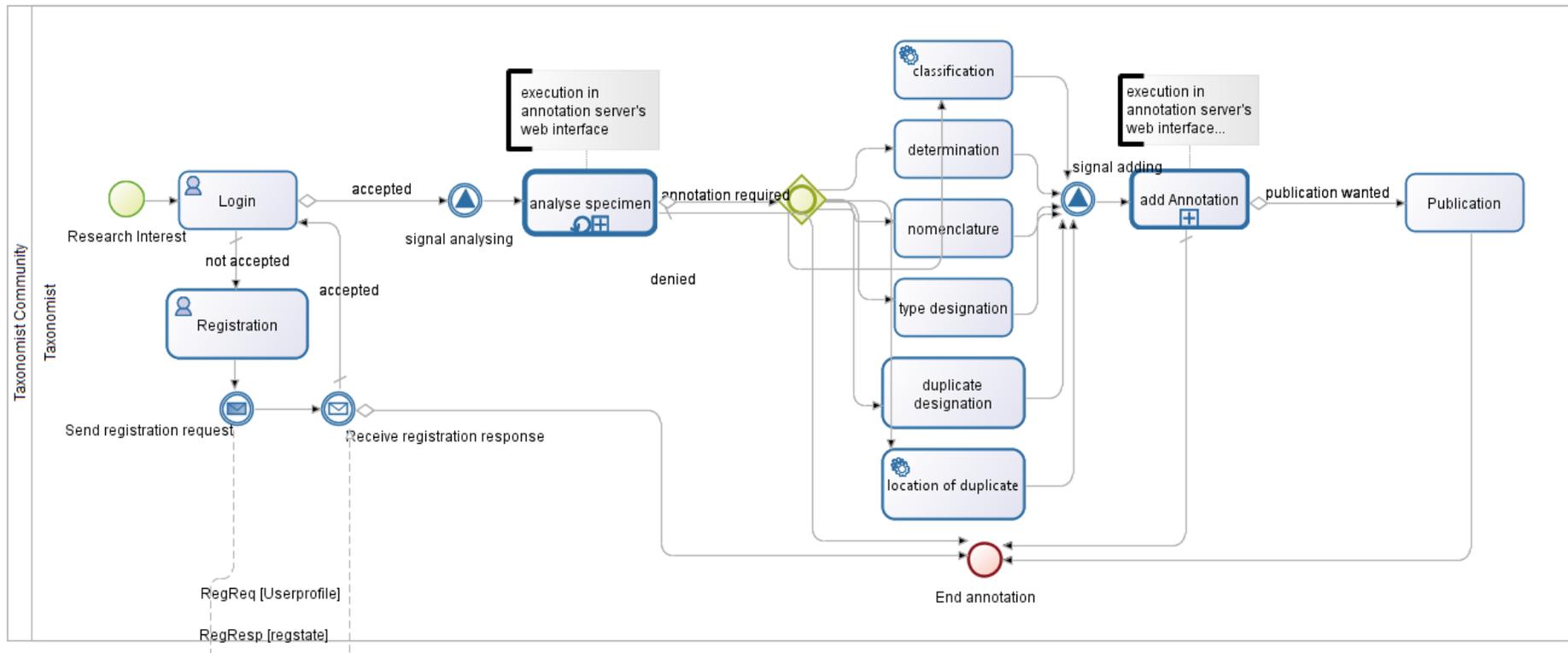


Figure 2: Annotation workflow in the BioCASE data portal (part 1)

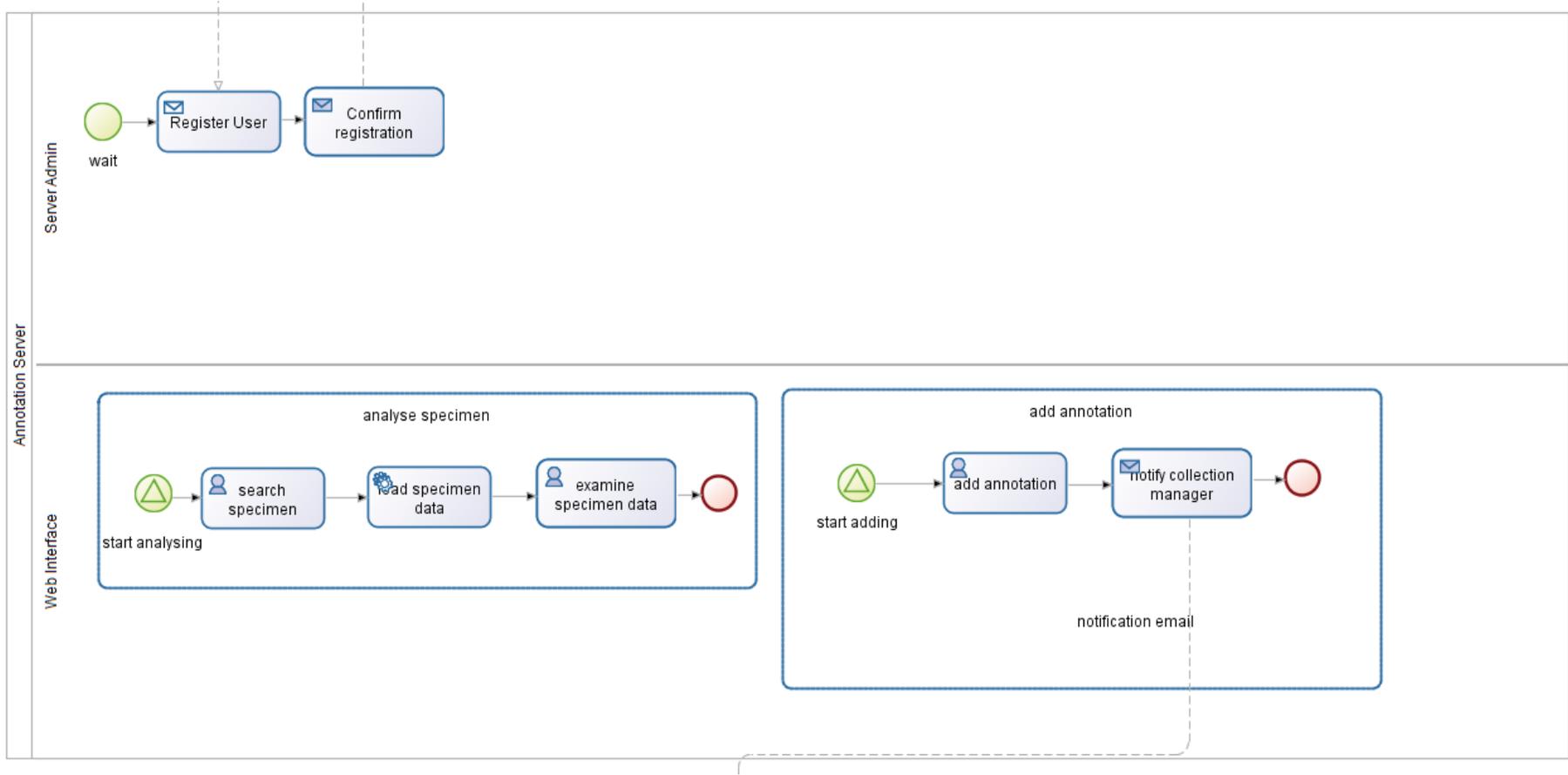


Figure 3: Annotation workflow in the BioCASE data portal (part 2)

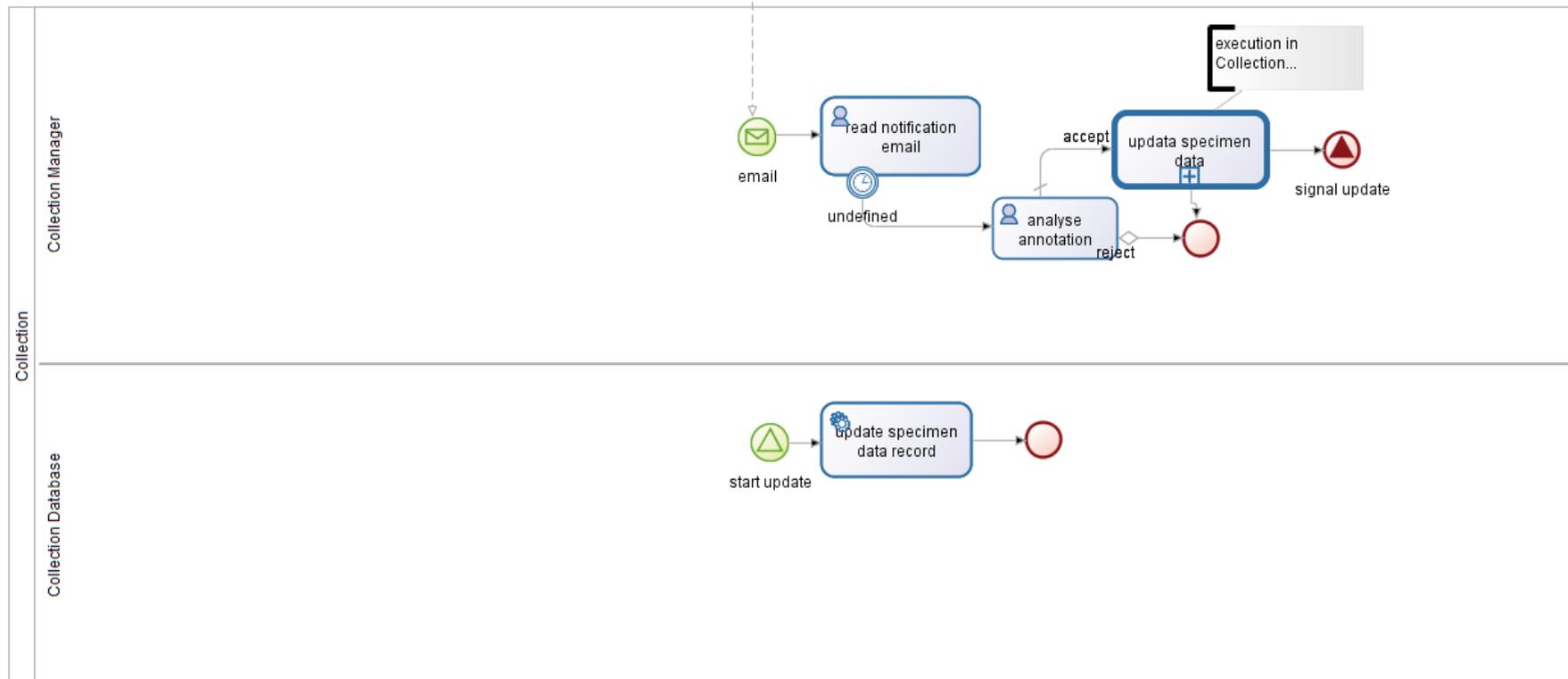


Figure 4: Annotation workflow in the BioCASE data portal (part 3)

JSTOR Plant Sciences

The JSTOR Annotation Mechanism

Digital imaging is carried out by several projects. This workflow describes the main tasks of the GPI (Global Plant Initiative) project „Digitizing of the Herbarium Willdenow at Berlin“. It shows the basic interaction with the JSTOR Plant Science service and how annotations (comments) that are entered at JSTOR will be (re)integrated.

The aim of the digitalisation project is to integrate several thousands of formerly not electronically captured specimen sheets into the Herbarium database. Therefore, specimen sheets have to be prepared one by one in order to get digitalised by special digital photographic equipment. After a picture has been taken, the resulting image has to be checked for quality. If the image quality is satisfactory, the image will be made accessible through an interface to a specific database for digitalisation purposes.

Next, any written or printed information on the sheets will be interpreted and entered as metadata for the dataset of the given specimen within the digitalisation database. Thereby, annotations that have been made on the specimen sheet are added to the comment field within the database form. Note that the comment field may also contain other information like the name of the specific herbarium (e.g. "Herb. Leimbach"), the number of sheets (Sheet 1/2), or others, and is not exclusively used for taxonomic annotations.

Now the content of the digitalisation database will regularly be uploaded and updated to JSTOR Plant Science service. Therefore, at the JSTOR platform, two different account types exist. One type is for institutional users in order to upload and maintain their datasets. The other type is for users wanting to store comments on given specimen datasets at JSTOR (i.e. annotate).

If a JSTOR user adds a comment to a given dataset, the uploading institution will be notified by email. Although the JSTOR platform permits to update specimen data directly, usually data update will be done by BGBM digitalisation team into the digitalisation database. Updating the JSTOR entries will occur on the next periodical upload cycle of the digitalisation database to JSTOR. Only if there is a severe mistake in the data the update will be done directly on the platform. After a comment/an annotation has been integrated in the dataset by the data provider the annotation is recorded in the "notes" field. The annotation also remains visible in the "Comment" section on the JSTOR platform. Direct communication between the digitalisation team and annotators will also rarely occur, but has to be done by exchanging private emails.

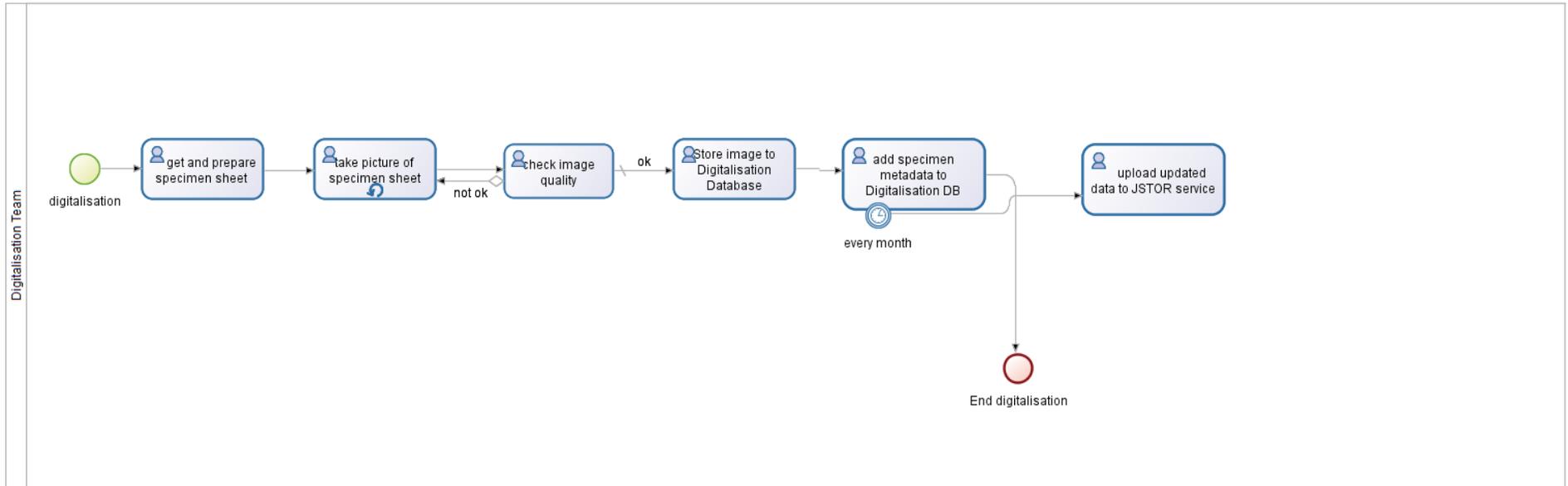


Figure 5: Workflow in the JSTOR platform (part 1)

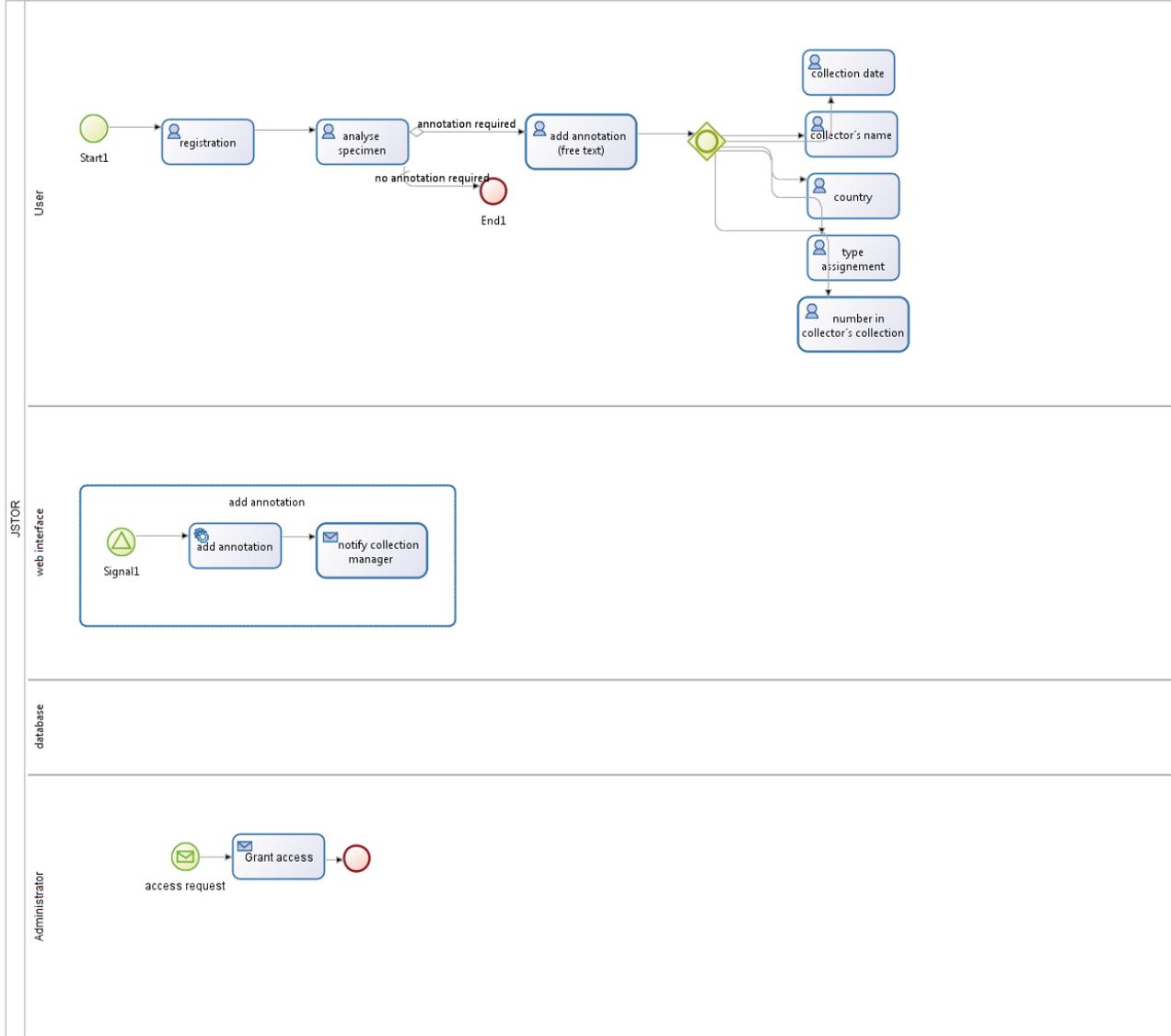


Figure 6: Workflow in the JSTOR platform (part 2)

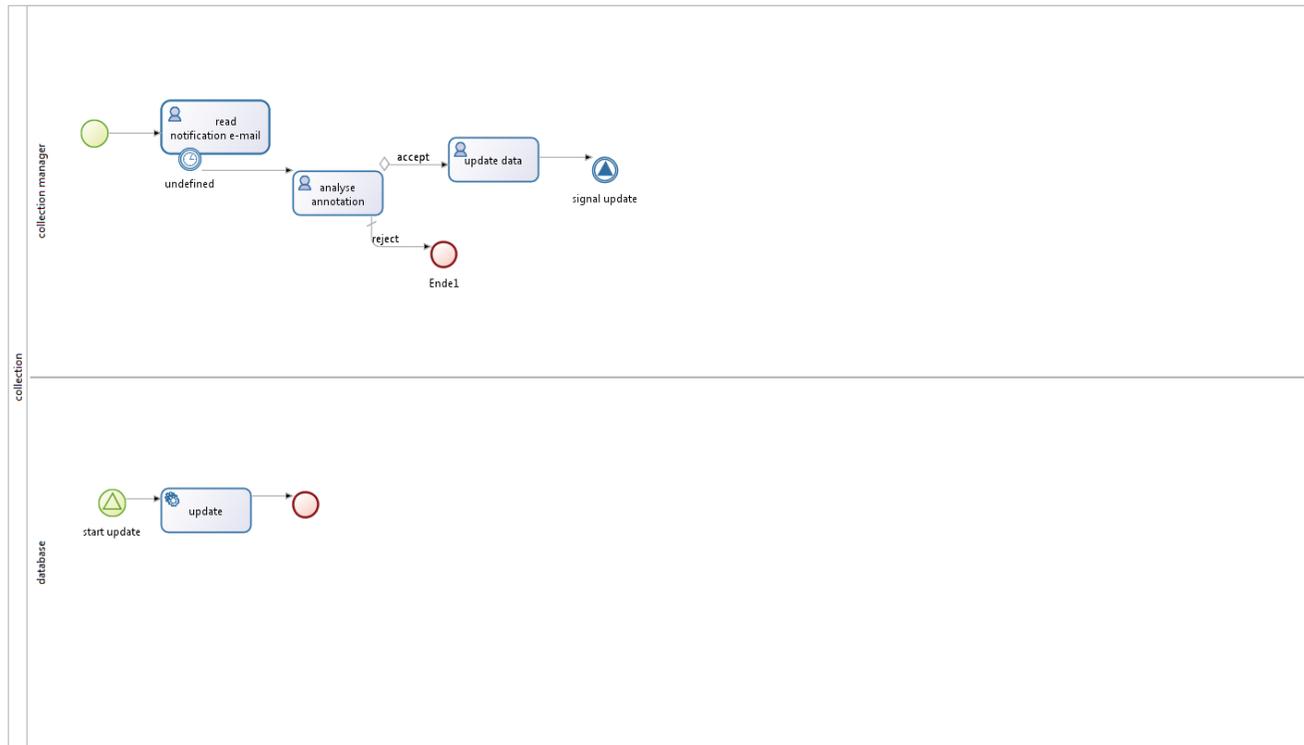


Figure 7: Workflow in the JSTOR platform (part 3)

The JSTOR Annotation import mechanism

For deliverable 3.5 of NA3, a prototypic “reverse wrapper” has been implemented. That term refers to the inverse data flow as compared to a “wrapper”, a software package used to feed data from primary biodiversity databases to data networks such as BioCASE or GBIF (Global Biodiversity Information Facility). An example for such wrapper software is the BioCASE provider Software⁷, which has been implemented by the BioCASE project and is still maintained by SYNTHESYS.

In the workflow diagram in figure 7, this prototype is located in the upper part (collection manager). Even though it doesn't change fundamentally, it changes in two ways:

1. The action “read notification e-mail” becomes “view annotation in database front end”, and
2. The “update data” action is eased for the collection manager by offering copy & paste functionalities, since the annotation will have been split up by the annotation loader into its components and can be easily copy & pasted between database forms.

The task of the prototype developed was to feed the annotations created by users on the JSTOR website back into the herbarium database. Because an immediate update of the herbarium database is neither desired nor technically trivial, they will be written into a separate table.

In a first step, the Annotation Loader will be run automatically every night and extract the annotations from the annotation emails sent by JSTOR. They will be parsed and broken down into their components, which will be written to a temporary annotation table in the herbarium database. This is implemented in Python⁸, a platform-independent scripting language available for most operating systems. Therefore, the annotation loader can be ported to any platform supported by Python.

In a second step, the Access front-end for the herbarium database was extended to display annotations if a specimen record is edited. The curator has the chance to browse through the annotations, confirm or reject them, and update the respective fields in the herbarium catalogue.

Filtered Push

The Filtered Push project is aimed at producing a system for improving the fitness for purpose of distributed data through analysis, annotation, and human review of data quality annotations. The software implemented will allow data providers and consumers to define potential errors in data, develop metrics for those errors, analyze distributed data to detect potential errors, and close the quality management cycle by providing a network architecture to move assertions about data quality such as corrections back to the curators of the original distributed data sets. The workflow is as follows.

Before a taxonomist can examine specimen datasets, he may have to make several queries into the Filtered Push network in order to get the desired datasets. These queries may include to request an inventory of available datasets or to find datasets according to search filter to be defined.

⁷ http://www.biocase.org/products/provider_software

⁸ <http://www.python.org/>

If the taxonomist detects a deficiency in one or more examined datasets, he may decide to make an annotation. The annotation may comprehend the following annotation types:

- Correction of data,
- Adding new information to the dataset, or
- Making a new determination of the given specimen.

Next, the annotation will be injected into the Filtered Push network via an annotation message. Therefore, the injecting client (either human taxonomists or software agents like quality checking modules) must provide an authentication token enabling the identification of the annotating user to the Filtered Push network.

While the taxonomist or agent has to identify the annotated dataset by a GUID (Global Unique Identifier, through software) or a Darwin Core triplet ("Institutional acronym:Collection code:Catalogue number"), the network stores the annotation and forwards it to a user who is collection manager of the collection database holding the authoritative record for the specimen or observation being annotated. For that, the collection database must be able to map annotations to or from its local database representation into a structured data format (e.g. ABCD, Darwin Core).

Now, the collection manager may accept, reject or ignore the annotation. In case of acceptance, the annotated data must be transformed into the local database schema before being stored into the authoritative local database. In case of rejection, no change is made to the authoritative local database. Both cases result in injecting either a notice of acceptance or rejection back into the network. The ladder will also retain knowledge of the annotation and its status. In case of ignorance, no further action happens until the collection manager decides to either accept or ignore the annotation.

In order to support collection managers in their decision processes, they must be enabled to filter out those annotations that are of potential interest to the datasets they curate. Also, the system must enable them to select which parts of which annotations to accept into their local databases.

Furthermore, the system must enable software agents to filter specific annotation types for potential (semi-)automatic processing.

Acknowledgement

The author is indebted to Okka Tschöpe and Lutz Suhrbier of the AnnoSys project for their support in compiling this report.

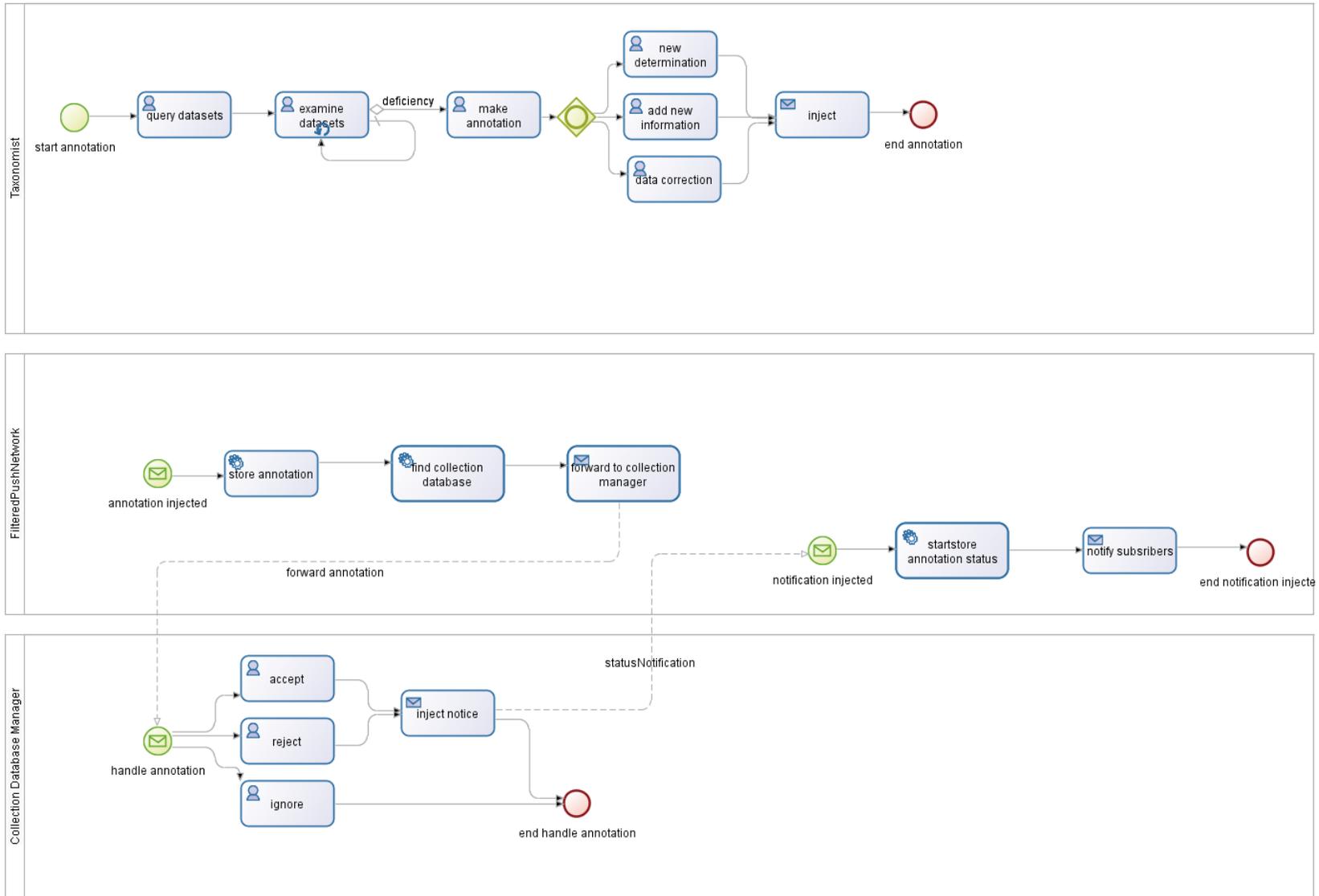


Abbildung 8: Filtered Push Workflow

