Mobilization of data from historical literature through markup, data extraction and re-publishing:

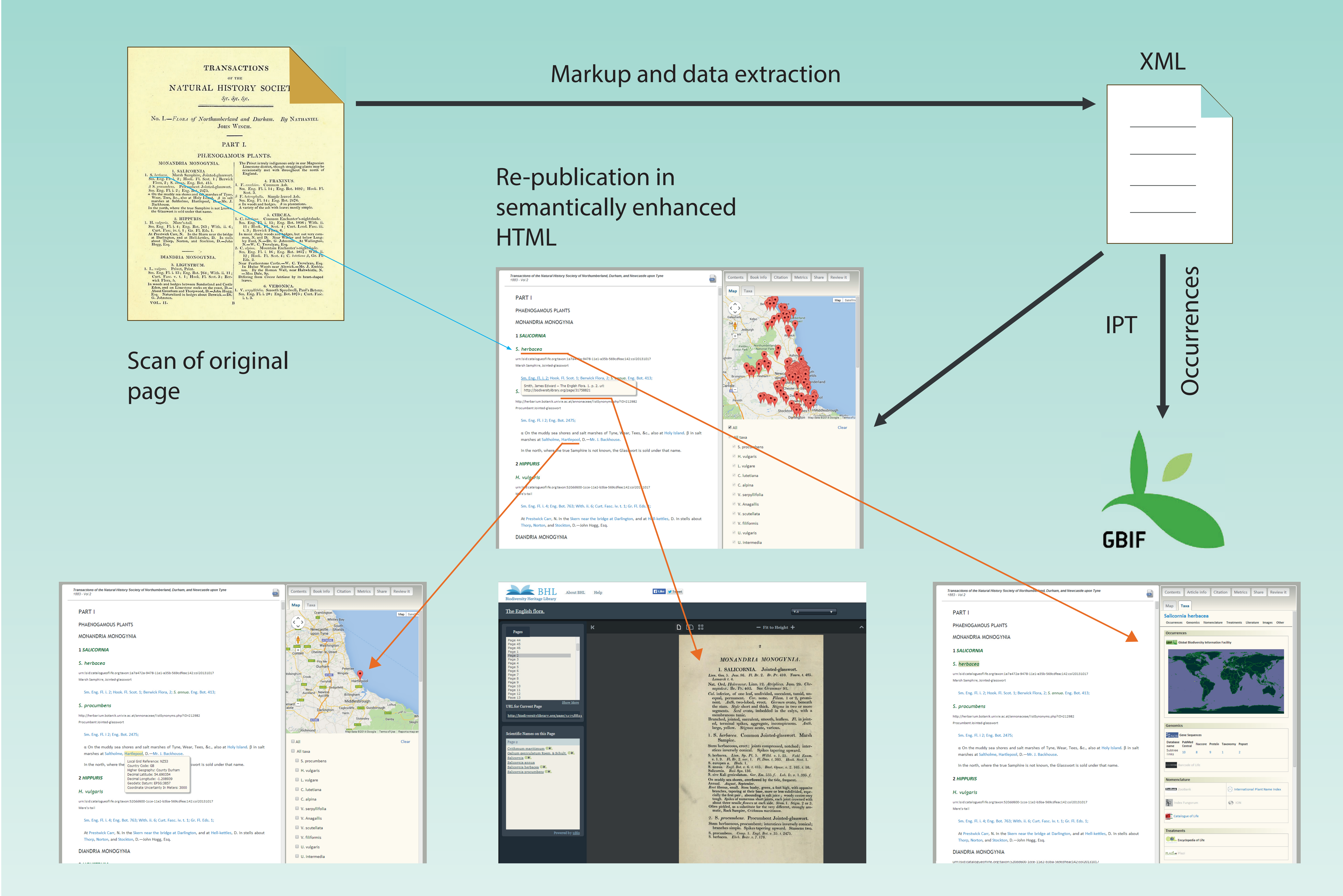
The Flora of Northumberland and Durham (1838) starts a new life!

Quentin Groom[1], Teodor Georgiev[2], Jordan Biserkov[2],
Pavel Stoev[2], Donat Agosti[3], Lyubomir Penev[2]

[1] Agentschap Plantentuin Meise, Meise, Belgium
[2] Pensoft Publishers, Sofia, Bulgaria
[3] Plazi, Bern, Switzerland

A huge amount of biodiversity data have been accumulated and published in hundreds of millions of pages of legacy literature. The process of reviving these data to be collated and re-used for the purposes of historical dynamics of biota and prognostic modelling for the future include three important steps:

- Markup and extraction of data, in this case occurrence records of plants published in the historical Flora of Northumberland and Durham (1838) and generation of an XML file of the content
- Putting data in a structured file and upload it onto GBIF through the Integrated Publishing Toolkit (IPT)
- Re-publishing of the historical flora from the XML into semantically enriched open access HTML version to facilitate readability and re-usability by humans



Markup and data extraction

XML

Re-publication in semantically enhanced HTML

Scan of original page

IPT

Occurrences

The text of the Transactions of the Natural History Society of Northumberland, Durham, and Newcastle-upon-Tyne was loaded in to Wikisource from the Internet Archive. The OCR text of the flora was corrected within Wikisource manually. Once the text was corrected a semi-automated process was used for markup. Wherever, possible formatting and punctuation were used to guide custom scripts in the placement of XML tags. Initially, main elements were identified, such as headings and treatment boundaries, then other elements were identified either manually, or using scripts. Progressively finer-grained markup was achieved. Each new element being validated against a custom XML schema to ensure valid XML. Some elements, such as the names of locations, had no standard format and were marked up by manually. However, as location names are often reused regular expressions were used to rapidly markup duplicates. Once the initial markup of elements was completed, the XML attributes for each tag where either added manually or using regular expressions.

The HTML version of the re-published flora offers several enhancement to the historical text, such as linking to external sources, mapping of the species, etc.

The establishment of the workflow was supported by the EU-funded project EU BON, grant agreement No 308454.

Literature

Winch, N.J. (1838) Flora of Northumberland and Durham. Transactions of the Natural History Society of Northumberland, Durham, and Newcastle-upon-Tyne. 2: 1–149.

PENSOFT®   Botanic Garden Meise   PLAZI taking care of freedom