Honest Broker for BioInformatics Technology (HoBBIT)

Luke Geneslaw[1,2,3] • Jennifer Samboy[1,2,3] • Vijay Yarlagadda[1,2,3] • Evangelos Stamelos[1,2] • Thomas Fuchs[1,2,3,4]

[1] Memorial Sloan Kettering Cancer Center • [2] Department of Pathology • [3] The Warren Alpert Center for Digital and Computational Pathology • [4] Weill Cornell Graduate School of Medical Sciences

What is HoBBIT?

HoBBIT is a database and pipeline for storing, de-identifying, and curating computational pathology research datasets. These datasets contain 2 types of data:

1. Discrete Pathology Report Data
2. Digital Pathology Images

What are HoBBIT's Use Cases?

Cohort Querying Examples:

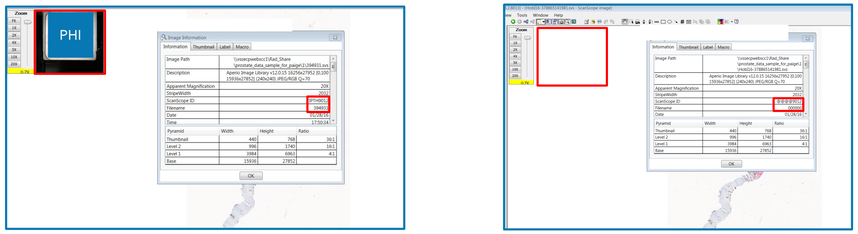
- How many images/cases are available that match a certain criteria? (i.e. specific synoptic fields selected)
- How many images are available from a given list of cases?

De-identification & Transfer:

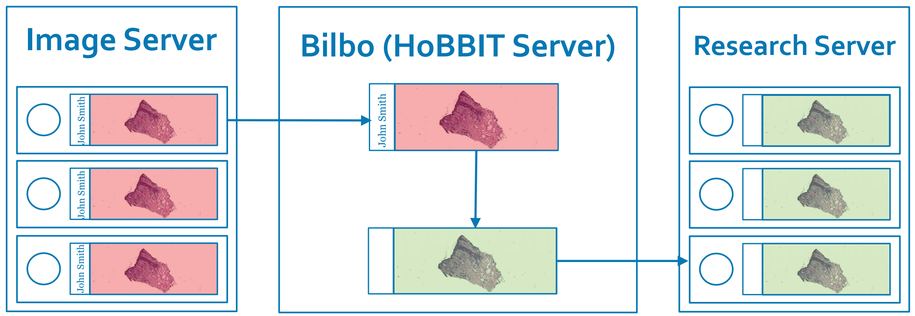
- Creates a de-identified copy of digital images on demand
- De-identifies and stores all pathology reports associated with digital images

Image De-identification and Transfer

Digital pathology image files contain patient identifiers in the image label and metadata. Each image is over 500MB, so it is not feasible to de-identify the entire MSK Archive. HoBBIT de-identifies images on demand and transfers them to a research server.



Screenshot of a pathology image before and after de-identification, which removes the label (top-left red box) and some metadata.



De-identification and transfer of digital pathology images via the Bilbo server. HoBBIT copies images from the clinical server and places them in a designated research server.

Pathology Report De-identification

HoBBIT extracts discrete pathology report fields from Copath and stores them in identified form. Some of these fields contain patient identifiers, so these fields undergo de-identification.

Types of De-identification:

- Date Truncation – Ex: Accession Date is truncated to Accession Year
- ID Creation – Ex: Each MRN is assigned a unique Patient HID
- Text Redaction – Ex: Final Diagnosis is redacted to remove identifiers

Pathology Report Fields in Moria and their De-identified Counterparts (Only bold fields require de-identification)

Accession #	**Case HID**
Accession Date	**Accession Year**
Specimen Class	*Specimen Class*
Part Type	*Part Type*
Part Instance	*Part Instance*
Part Description	**De-identified Part Description**
Block Instance	*Block Instance*
Block Designator Label	**De-identified Block Designator Label**
MRN	**Patient HID**
Aperio Slide Image ID	**Image HID**
Scan Date	**Scan Year**
Stain	*Stain*
Synoptic Data	*Synoptic Data*
Final Diagnosis	**De-identified Final Diagnosis**

How Does HoBBIT Get Data?

Moria (HoBBIT db) extracts and stores pathology report data & image metadata according to the following pipeline: