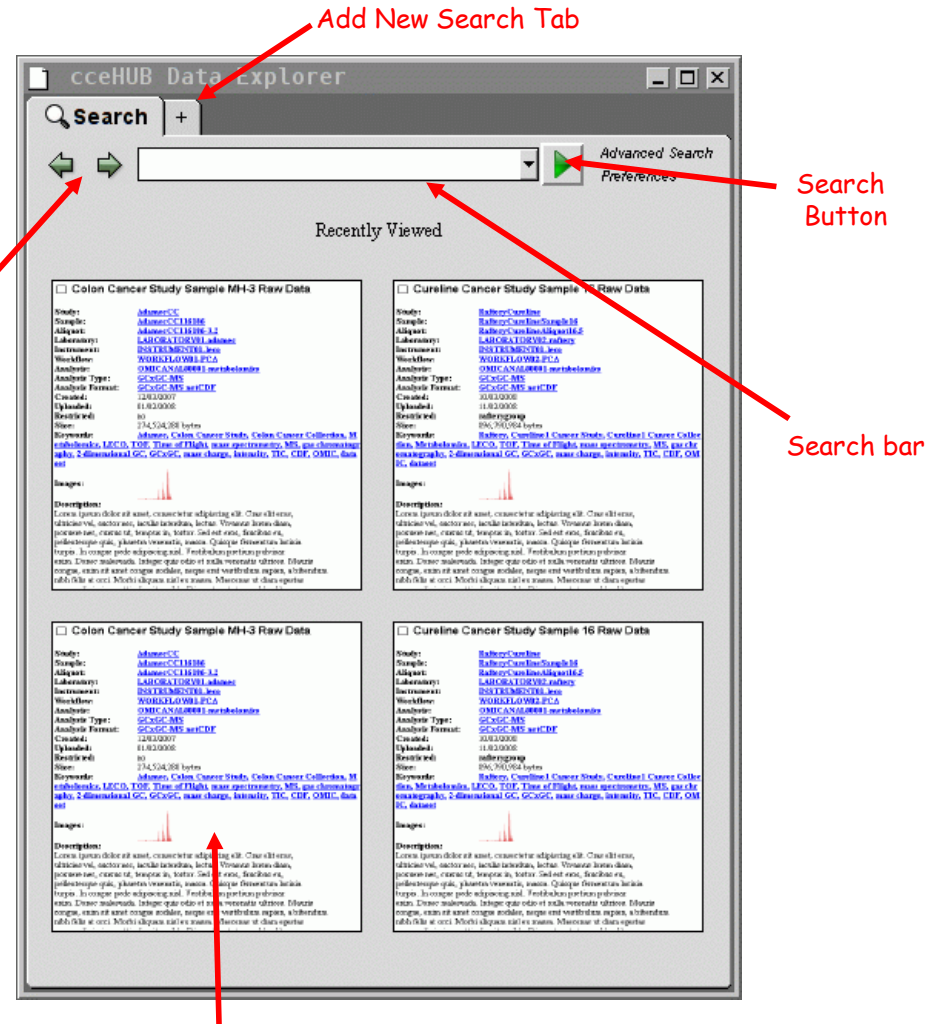


# cceHUB Data Explorer

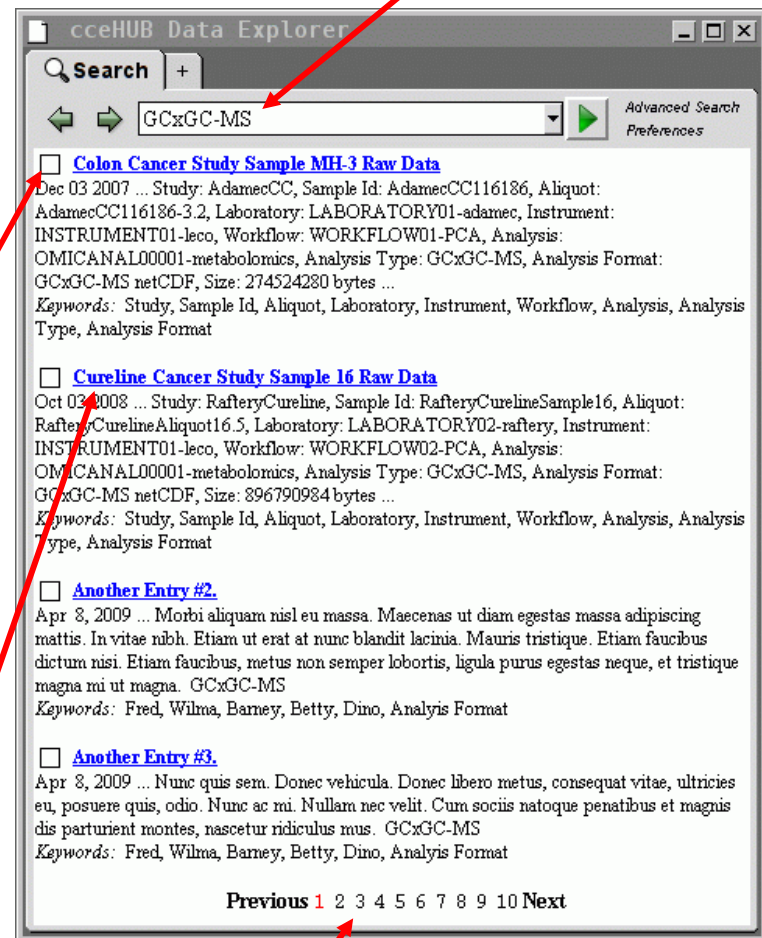
- Problem: Lots of data
  - » Samples, new tool results, etc.
  - » Very large datasets, MBs or GBs.
  - » How to easily select data/results of interest.
- Browser-like data explorer
  - » Search metadata via keywords (tags) and free text.
  - » Leverages metadata and keyword processors.
- Two different applications
  - » Data selector from inside tool (analysis, visualization, etc).
  - » Generalized data browser.



# cceHUB Data Explorer

- Google-like search mechanism
  - » Can be context free: “GCxGC-MS” vs. “AnalysisType:GCxGC-MS”
  - » Check boxes to select datasets.
- Search results are ranked
  - » Keyword, full/partial match, etc.
  - » Results also scored by user/group popularity.
- Efficiently queries database
  - » Returns IDs of matches, sorted by score.
  - » Full data content requested on demand.

Query matches keyword and text



Checkbox

Clicking on title displays content page.

View page worth of entries at a time.

# cceHUB Data Explorer

- Entry page contains
  - » Description of dataset.
  - » Check box to select entry.
- But also...
  - » Links to category pages. example: “Instrument”.
  - » Links to tools.
  - » Images (graphs) from tool results.
- Keyword links perform search.
- Leverage Rappture to provide easy connection with tools.

The screenshot shows the 'cceHUB Data Explorer' window. At the top, there is a search bar with 'GCxGC-MS' entered. Below it, a dropdown menu shows 'GCxGC-MS' and a green play button icon. To the right of the dropdown are links for 'Advanced Search' and 'Preferences'. The main content area displays a dataset entry:
 

- Colon Cancer Study Sample MH-3 Raw Data
- Study:** [AdamecCC](#)
- Sample:** [AdamecCC116186](#)
- Aliquot:** [AdamecCC116186-3.2](#)
- Laboratory:** [LABORATORY01-adamec](#)
- Instrument:** [INSTRUMENT01-leco](#)
- Workflow:** [WORKFLOW01-PCA](#)
- Analysis:** [OMICANAL00001-metabolomics](#)
- Analysis Type:** [GCxGC-MS](#)
- Analysis Format:** [GCxGC-MS netCDF](#)
- Created:** 12/03/2007
- Uploaded:** 01/02/2008
- Restricted:** no
- Size:** 274,524,280 bytes
- Keywords:** [Adamec](#), [Colon Cancer Study](#), [Colon Cancer Collection](#), [Metabolomics](#), [LECO](#), [TOF](#), [Time of Flight](#), [mass spectrometry](#), [MS](#), [gas chromatography](#), [2-dimensional GC](#), [GCxGC](#), [mass charge](#), [intensity](#), [TIC](#), [CDF](#), [OMIC](#), [dataset](#)
- Images:**
- Description:** Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras elit eros, ultricies vel, auctor nec, iaculis interdum, lectus. Vivamus lorem diam, posuere nec, cursus ut, tempus in, tortor. Sed est eros, faucibus eu, pellentesque quis, pharetra venenatis, massa. Quisque fermentum lacinia turpis. In congue pede adipiscing nisl. Vestibulum pretium pulvinar enim. Donec malesuada. Integer quis odio et nulla venenatis ultrices. Mauris congue, enim sit amet congue sodales, neque erat vestibulum sapien, a bibendum nisl felis at orci. Morbi aliquam nisl eu massa. Maecenas ut diam egestas massa adipiscing mattis. In vitae nisl. Etiam ut erat at nunc blandit lacinia. Mauri's tristique. Etiam faucibus dictum nisi. Etiam faucibus, metus

Check box

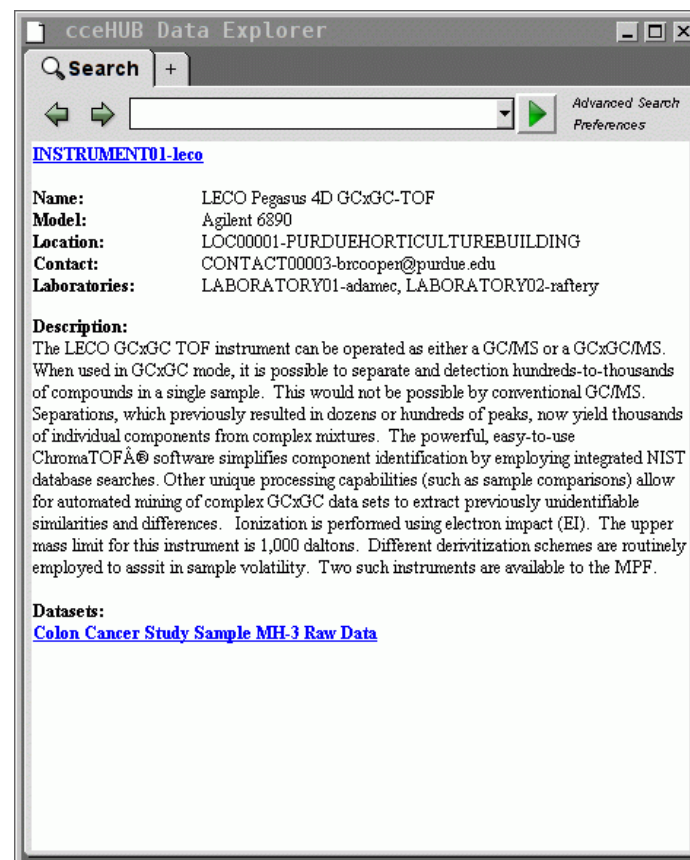
Links to more content.

Keyword links

Link to tool screenshot.

# cceHUB Data Explorer

- Content pages
  - » Display more general information.
  - » Describe how data was collected/processed.
  - » Link to similar datasets.
  - » Simple HTML (generated by backend process).



The screenshot shows a web browser window titled "cceHUB Data Explorer". At the top, there is a search bar with a magnifying glass icon and a plus sign. Below the search bar are navigation arrows and a dropdown menu. To the right of the search bar are links for "Advanced Search" and "Preferences". The main content area displays the following information:

**[INSTRUMENT01-leco](#)**

**Name:** LECO Pegasus 4D GCxGC-TOF  
**Model:** Agilent 6890  
**Location:** LOCD0001-PURDUEHORTICULTUREBUILDING  
**Contact:** CONTACT00003-brcooper@purdue.edu  
**Laboratories:** LABORATORY01-adamec, LABORATORY02-raftery

**Description:**  
The LECO GCxGC TOF instrument can be operated as either a GC/MS or a GCxGC/MS. When used in GCxGC mode, it is possible to separate and detection hundreds-to-thousands of compounds in a single sample. This would not be possible by conventional GC/MS. Separations, which previously resulted in dozens or hundreds of peaks, now yield thousands of individual components from complex mixtures. The powerful, easy-to-use ChromaTOF<sup>®</sup> software simplifies component identification by employing integrated NIST database searches. Other unique processing capabilities (such as sample comparisons) allow for automated mining of complex GCxGC data sets to extract previously unidentifiable similarities and differences. Ionization is performed using electron impact (EI). The upper mass limit for this instrument is 1,000 daltons. Different derivitization schemes are routinely employed to assit in sample volatility. Two such instruments are available to the MPF.

**Datasets:**  
[Colon Cancer Study Sample MH-3 Raw Data](#)

# cceHUB Data Explorer

- Google Model
  - » Ranked results (web2.0).
  - » Unstructured data searches.
- Structured data vs. unstructured data
  - » Structured search: query reflects knowledge of database schema.
    - ✓ Good for programs. Range searches. When you know what you want.
  - » Unstructured search: structure imposed on-the-fly as searched.
    - ✓ When you don't know exactly what you are searching for.
  - » Most important point is how easy to find/access your data.
- Data explorer uses both methods.
  - » Displays a table of search results.
  - » Table entries can be sorted and narrowed.
  - » Good for range searches. “*Show entries where the peak retention time is between...*”

# cceHUB Data Explorer

- Future plans.
  - » Implement table range searches.
  - » Be able to invoke tools from within data explorer.
  - » Connect to cceHUB database.
  - » Connect to keyword and metadata processors.
  - » Score pages by popularity.