Shared Print
Gold Rush Library Content Comparison

George Machovec, Executive Director
Colorado Alliance of Research Libraries

January 8, 2016
Print Archive Network (PAN)
ALA MidWinter
Boston, MA

george@coalliance.org

Colorado Alliance of Research Libraries

- Incorporated in 1981
- 14 member libraries (13 academic, 1 public) – one out of state library (University of Wyoming)
- Programs include
  - E-resource licensing – 250 contracts (>$14 million)
  - Prospector union catalog – 44 libraries, 13.5+ million unique MARC records
  - Gold Rush (ERMS, link resolver, content comparison)
  - Shared Print (launching 2015)
- A history of innovation and software development

Why another shared print program?

- Regional focus
- Running out of room in our libraries and storage facilities
  - Several renovation projects underway
- Strong resource sharing network
  - Prospector union catalog
- Ever greater reliance on ebooks
- Several storage facilities in the state

Key Features

- Distributed
  - No attempt to build a single journal run
  - No central shared storage facility shared by all although a shared storage facility for 4 CU campuses exists
  - Each library maintains holdings
- Voluntary
  - Libraries can participate as much or as little as they wish
  - No one is ever forced to keep or discard a volume
- Flexible
  - Can expand as needed
  - Allows libraries to participate in other programs

Memorandum of Understanding

- Signed by participating institutions
  - But every institution benefits
  - Can bring in non-Alliance partners as needed
- Establishes the Alliance Shared Print Trust
  - Distributed print repository
  - Participants agree to retain materials on behalf of the group and disclose retention decisions
- Provides a framework for specific projects, which can be established as needed
- 25-year commitment, reviewed every five years

Shared Print Program

- Alliance Shared Print Trust Agreements
  - Broad agreement (MOU) signed by participating libraries
  - Circulating monographs policy, Serials policy, Disclosure policy
- An analysis tool was needed that was affordable and flexible
- Commercial tools (Intota Assessment, OCLC Collection Evaluation, SCS Greenglass for Groups) are excellent services but were too expensive for member libraries

Comparison Tool Selected Characteristics

- Must be scalable to work from the smallest to largest libraries
- Must work in real-time
- Must support comparing any combination of libraries (1-1, 1-many, many-many)
- Must support export of MARC, XML and delimited (Excel)
- Must show what's unique and overlapping, including visualizations
- Must be able to have searches or facets to work with any fields in a MARC record
- Matching algorithm must be made from elements in a MARC record but **NOT** depend on OCLC #, ISBN, ISSN
- Must be easy-to-use!

Elements in Match Key

- **Title**
  - 245 $a $b
- **General Media Description**
  - 245 $h
- **Type of**
  - '_' Leader
- **Title Part**
  - 245 $p
- **Title Number**
  - 245 $n
- **Publication Year**
  - 260 or 264 $c
- **Pagination**
  - 300 $a
- **Edition Statement**
  - 250 $a
- **Publisher Name**
  - 260 or 264 $b
- Additional programming for cartographic and sudocs

Technology

- Linux (CentOS)
- “Play” web application framework
- SOLR (developing SOLR RDF) – no SQL
- Open source charting software
- All virtualized on multiple VMs
- Google MARC ingest (from the Google Books Project)
- Must be scalable to hundreds of millions of records working in real-time

Loading and updating

- System is ILS agnostic
- Records can be done as often as you want but no more than once per day
- Suggested that full database updates be done monthly (unless a big change occurs and then it can be done on-demand)
  - No deletes, at the present, so full updates are needed when many records are removed from your local catalog
- SFTP (secure FTP) is used to deposit records in either a “full” or “updates” directory

Source of MARC records

- Prefer a direct export from your ILS rather than a union catalog export
- Challenges with union catalog (INN-Reach) metadata
  - Nobody contributes all MARC records
  - The master record may not always be yours
  - Libraries prefer getting their own records back rather than a generic MARC record from the union catalog
  - A local export will eventually be need for item and/or circ data
- But the Prospector union catalog was perfect for development and scaling

https://grcompare.coalliance.org/prospector/noblChart/?utf8=✓&search_field=all_fields&q=&site_1=9cubp&site_2=9uwyp&compareAction=&startYear=1000&endYear=2016&resultCount

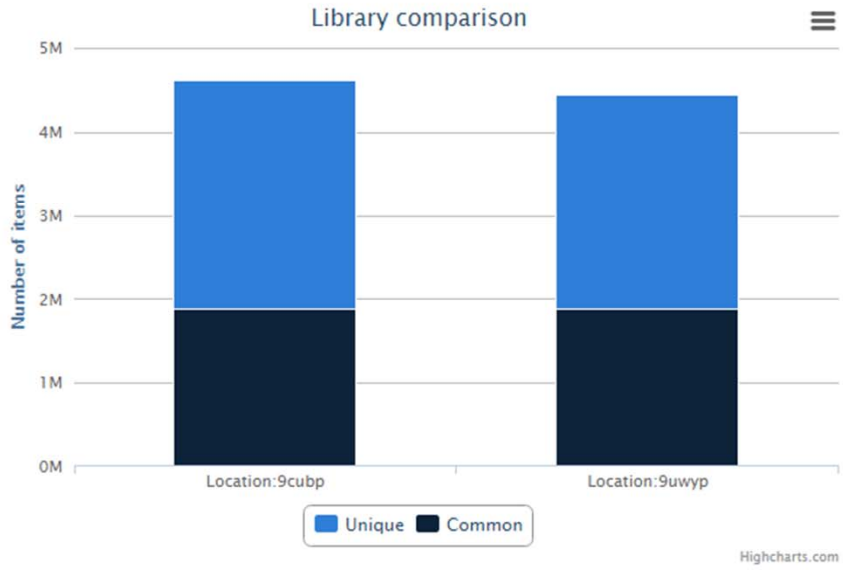
Gold Rush » Reports gsmachovec@gmail.com

All Fields | Search... | Search | CU Boulder (9cubp) | 1 selected | Site Compare

1884438 found! Displaying documents 1 to 25 Previous Next | Save search | Load search | Export Marc

Limit your search...

- Format
- Publication year
- Subject heading
- Language
- Call number
- Region
- Era
- Site code
- Local bib #



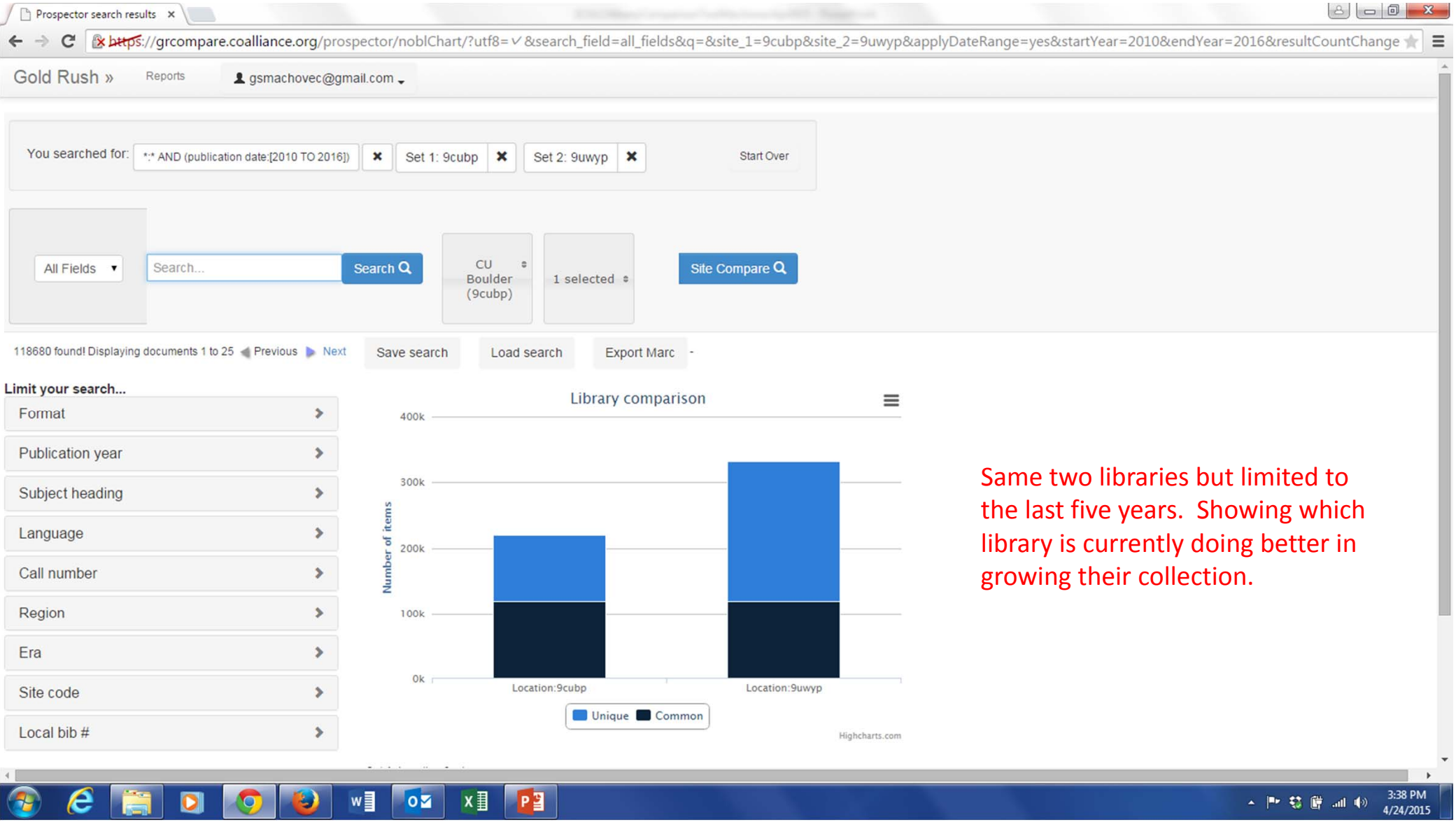
Set 1: Location:9cubp
has 2728018 unique , with 1884438 common and a total of 4612456 items.

Set 2: Location:9uwyp
has 2548625 unique , with 1884438 common and a total of 4433063 items.

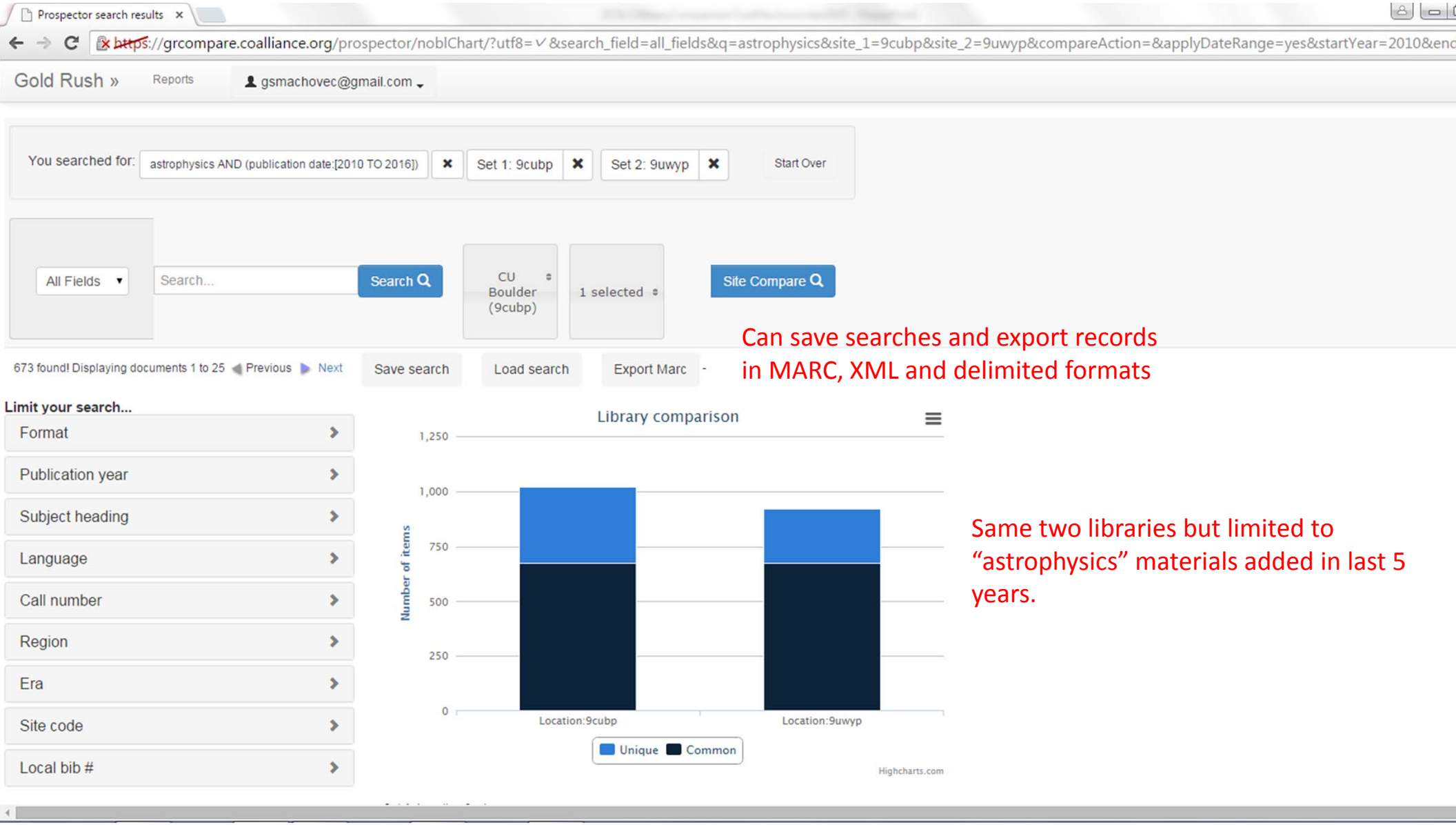
Run time: 8715 milliseconds

Simple comparison of two large academic libraries with no special limiting. Each with greater than 4 million records. University of Colorado vs. University of Wyoming

At this level mostly useful for bragging rights!

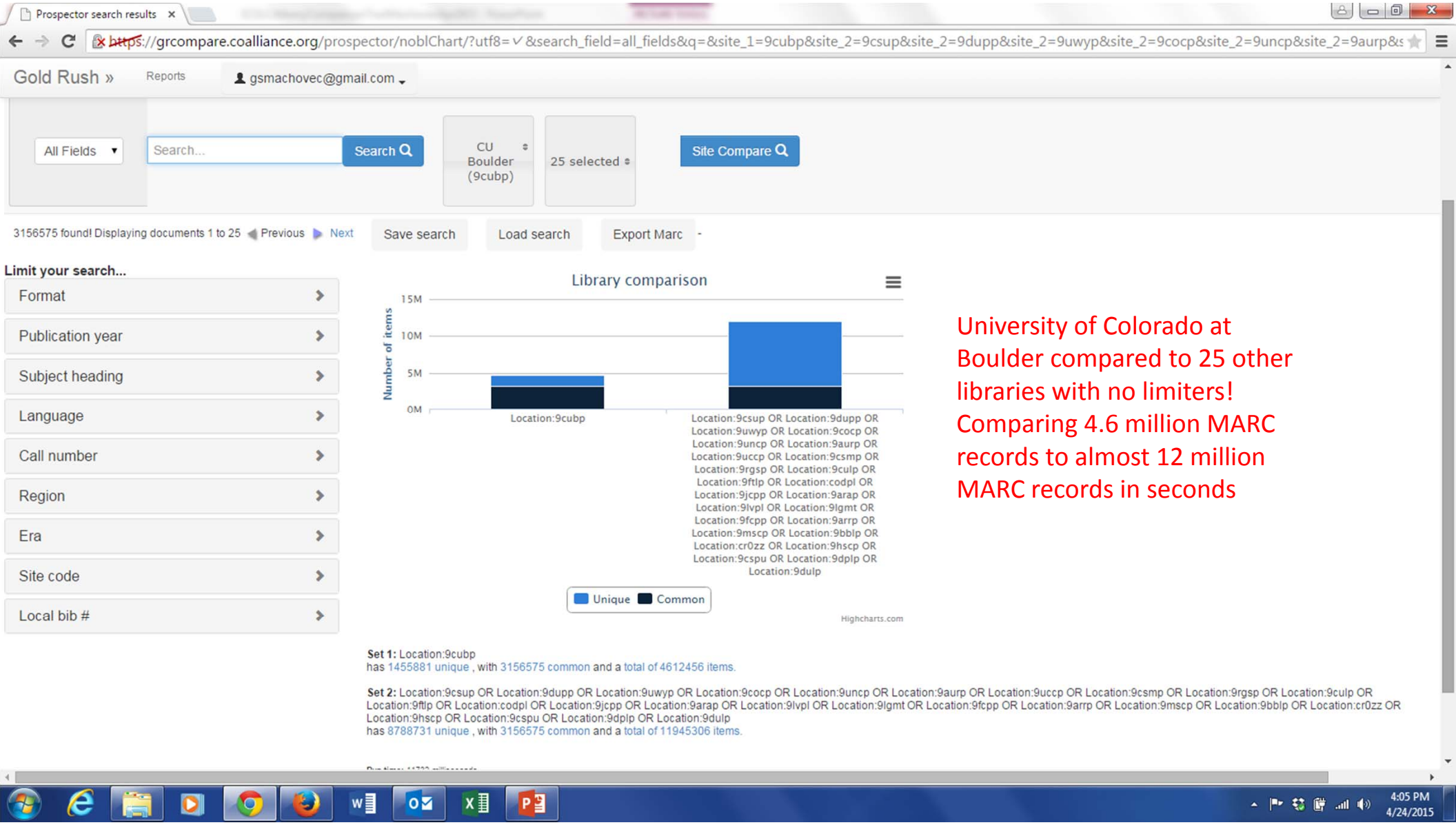


Same two libraries but limited to the last five years. Showing which library is currently doing better in growing their collection.



Can save searches and export records in MARC, XML and delimited formats

Same two libraries but limited to “astrophysics” materials added in last 5 years.



University of Colorado at Boulder compared to 25 other libraries with no limiters! Comparing 4.6 million MARC records to almost 12 million MARC records in seconds

Challenges Faced in the Project

- Making the user interface (UI) easy and intuitive
- Getting the matching algorithm right
  - Too tight we end up with too many orphans that should have matched
  - Too loose we end up with false merges
  - Sometimes libraries want e-resources to match with print and sometimes they don’t
  - This will likely keep changing as it will continue to be refined
- Making it all happen in real-time and scalable to any library size (accomplished)

Work Yet to be Done

- Interest in adding additional datasets such as HathiTrust public domain items, commercial ebook sets, etc.
- Adding branch level metadata for doing comparisons within a library system (particularly important for public libraries)
- Incorporating circ data
- Looking at overlap for “how many owning libraries” (right now it’s binary – is something unique or owned by more than one)
- Adding new facets as requested by users
- Conspectus type analysis – call number range analysis
- **Not** doing gap analysis for serials (although serials loaded for title-level comparisons)

Sample Use Cases

- Shared print programs
  - What to put in storage
  - What to weed
  - Getting a set of unique records and adding a 583 retention note
- Adding a new program on campus and want to compare holdings with an established institution with the same program
- Compare a consideration pool (e.g. weeding or storage) to others
- Want to compare branches in a public library system
- Comparing collection with commercial datasets or other catalogs (e.g. HathiTrust, CRL)
- Quick exports for participation in other cooperative programs
- Bragging or complaining with your colleagues or administration!

What's Next for the System?

- Getting direct loads from all of the Prospector (union catalog) libraries to replace the union catalog records
- Adding features of interest to participants – we are in control!
- Gearing up the infrastructure
- Optimizing many aspects of the system
  - Comparison speeds
  - Loading speeds
  - Better fault tolerance for bad metadata

Next Steps

- Promote the use of the content comparison tool (we will make it available to others outside of our region for a small annual fee)
- Build on existing programs of study and institutional collection strengths
- Coordinate monograph approval plans and purchasing of print monographs – enough copies, but not too many
- Purchase at least one copy of a print monograph for selected publishers and/or subjects

Questions?

George Machovec
Executive Director
Colorado Alliance of Research Libraries
george@coalliance.org