HOW TO SURVIVE THE INSTALL & DATA NORMALIZATION TO GET PRETTY RESEARCH INFORMATION

DR. LAURA HALE AND MIGUEL PARAZ
SURVIVAL SKILL #1: BUILDING A GOOD TEAM

KNOW WHAT YOU WANT TO DO AND BUILD THE TEAM TO MAKE IT HAPPEN
WHAT IS YOUR GOAL IN USING WIKIBASE, THE QUERY ENGINE AND BLAZEGRAHP?

WHY DON’T OTHER SOLUTIONS WORK?
THREE OR MORE EXPERTS REQUIRED

Technical expert

Domain knowledge data expert

Query expert

By OpenClipart - OpenClipart, CC0, https://commons.wikimedia.org/w/index.php?curid=824285
SURVIVAL SKILL #2: ANTICIPATE & PLAN FOR TECHNICAL CHALLENGES

UNDOCUMENTED CHALLENGES NEED TO BE PLANNED FOR
THE TECHNICAL SIDE CAN BE ANNOYING!

THE RIGHT PEOPLE

- SPECIAL TECHNICAL EXPERTISE IS REQUIRED.
- LACK OF TECHNICAL EXPERTISE TO ASSIST NOVICE DATA FOLKS.
- PROJECTS NEED THE RIGHT TECHNICAL PEOPLE TO GET OFF THE GROUND.

TECHNICAL LIMITATIONS

- THE INFRASTRUCTURE REQUIRES SETTING UP THE CLOUD INFRASTRUCTURE, SERVERS AND APPLICATIONS BY HAND.
- BLAZEGRAF STORES RDF DATA IN MEMORY, WHICH IS EXPENSIVE AND LIMITS SCALABILITY.
- WIKIBASE EXPORT TO RDF USES HIGH CPU AND DATABASE RESOURCES, MAKING ITS WEB INTERFACE UNUSABLE DURING THIS TIME.
- NO MYSQL DISK SIZING GUIDELINES.

BLAZEGRAF

- AUTOMATING THE BLAZEGRAF QUERY ENGINE REBUILDING.
- BLAZEGRAF CRASH RECOVERY
...VERY ANNOYING!

**DOCUMENTATION**
- DIFFICULT TO FIND, MISSING OR WRONG!

**DATA IMPORT**
- WHAT SOLUTION IS THE BEST? WIKIDATA TOOLKIT EXAMPLES IS ONE OPTION
- YOU MAY NEED TO **DESIGN YOUR OWN**, LIKE PARASPORTS DID AT HTTPS://GITHUB.COM/MPARAZ/PARASPORTS-DATA-IMPORT

**COMMUNITY SUPPORT**
- CAN BE CHALLENGING.
- YOU MAY BE ON YOUR OWN.
IN YOUR QUERY ENGINE, TRY NOT TO BREAK THINGS...

TECHNICAL ASPECTS OF INSTALLATION EXTEND TO ISSUES THAT INCLUDE INSTALLING BLAZEGRAF AND THE QUERY ENGINE, AND THEN RUNNING QUERIES TO CHECK FOR ACCURACY. CROSS CHECKING THESE ON WIKIDATA'S QUERY SOMETIMES LED TO HANGING THE WMF'S BLAZEGRAF INSTALL.
SURVIVAL SKILL #3: STRUCTURE THE DATA CAREFULLY

DATA STRUCTURING IS KEY SURVIVAL SKILL
DATA STRUCTURE: PLAN IT!

• You need to plan data structure. It requires domain knowledge expertise.
• Data structuring using items and statements for Wikibase requires tons of planning.
• Plan early to avoid problems later.
PARASPORTS DATA STRUCTURE

MAJOR CONCEPTS FOR TOP LEVEL ITEMS

people
- role in sports
  - sportsperson
  - coach
- nationality
  - Canada
  - Australia
- disability
  - Vision impairment
  - Hearing impairment

events
- Dates held
  - Start: 2016 June 5
  - End: 2016 June 7
- Location
  - Madrid
- Sports on program
  - Goalball
  - Wheelchair basketball
  - Cp football

sports
- federation
  - International Paralympic Committee
- Eligible disabilities
  - amputees
  - Wheelchair users
- Variant od
  - Competitive swimming
DOMAIN KNOWLEDGE IS KEY TO DATA STRUCTURING

PARASPORT’S DATA STRUCTURE WAS BUILT AS A RESULT OF HAVING LARGE AMOUNTS OF DOMAIN KNOWLEDGE. IT REQUIRED A COMPLETELY SEPARATE SKILLSET FROM THE TECHNICAL ASPECTS.

IT REQUIRED ASKING QUESTIONS LIKE:

• WHAT ARE THE MOST IMPORTANT AND HIGH LEVEL CONCEPTS LIKELY TO GET THE MOST USE?
• WHAT TYPES OF ITEMS ARE MOST LIKELY TO BE ADDED? WHAT INFORMATION IS REQUIRED FOR THOSE?
• HOW AND WHERE SHOULD HIGH LEVEL TOPICS INTERSECT AND CONNECT?
• HOW CAN IMPORTANT DETAILS BE ADDED WHEN DEALING WITH EMBEDDED CONCEPTS LIKE THE NUMBER OF BASKETS MADE BY A SPECIFIC PLAYER IN A SPECIFIC GAME?
• WHAT SORT OF QUERIES WOULD OTHER EXPERTS LIKELY BE INTERESTED IN MAKING?
SURVIVAL SKILL #4: NORMALIZE DATA

LEARN HOW TO TAKE 15 SOURCES AND COMBINE THEM TO 1 ITEM.
DATA NORMALIZATION

MOST PEOPLE NEED TO COMBINE DATA FROM SEVERAL SOURCES. EXAMPLE SCREENCAPS ABOVE SHOW 3 SOURCES WITH VASTLY DIFFERENT TYPES OF DATA INPUT FOR SWIMMERS.

DATA NORMALIZATION OCCURS ALL THE TIME, WITH PLANNING STARTING WHEN YOU DO DATA STRUCTURING.
DATA NORMALIZATION

DATA PUT INTO SPREADSHEET

CONVERT TO FORMAT FOR UPLOAD

UPLOAD AND VERIFY

THIS IS THE MOST **TIME CONSUMING** PART FOR THE DOMAIN SUBJECT EXPERT.
SURVIVAL SKILL #5: LEARN TO WRITE QUERIES

LEARN HOW TO ASK QUESTIONS USING CODED LANGUAGE
QUERYING DATA

ALL THE SOFTWARE IS WORKING. DATA HAS BEEN UPLOADED AND CHECKED. TIME TO ENGAGE IN ANALYSIS!

WHERE TO START? FIRST, BY KNOWING ANY SPECIAL PREFIXES THAT MAY BE REQUIRED TO ALLOW QUERIES TO RUN ON A SPECIFIC INSTALL.
QUERYING DATA

NEXT, EXAMPLES SHOULD BE LOOKED AT, ON WIKIDATA, ON BLAZEGRAPH AND ON ANY OTHER WIKIBASE INSTALLS THAT CAN BE FOUND.

**Modelling** is the easiest way for most people to write queries. The more that is modelled, the easier it becomes to create workable queries, save them and build on them.

```
PREFIX ps: <http://> 
PREFIX pq: <http://>
#defaultView:LineChart
SELECT ?year (COUNT(? sport) AS ?count) (SAMPLE(? sportLabel) AS ? sportLabel) WHERE {
?item wdt:P203 ? date.
BIND(str(year(? date)) AS ?year)
FILTER((LANG(? sportLabel)) = "en")
FILTER(? date >= "1960-00-00T00:00:00Z"^^xsd:dateTime)
}
GROUP BY ? sportLabel ?year
HAVING (?count > 1)
```
PRETTY DATA WITH MEANINGFUL RESULTS

Like data structuring, querying is something that takes time to understand.

It may also require finding Wikimedia community members that specialize in writing queries and technical support that understands writing queries.

This is probably not the same person who installs the query engine.

Pretty data from ParaSports. Graph shows how vision impaired female swimmers outperformed their partially blind counterparts. It calls into question vision impaired classification as it relates to differences between classes and fairness in the system.
SURVIVAL SKILL #6: SEEKING ADDITIONAL HELP

THE ART OF ASKING FOR ASSISTANCE IS THE MOST IMPORTANT SURVIVAL SKILL FOR MANAGING A WIKIBASE INSTALL
MORE RESOURCES FOR YOUR OWN WIKIBASE INSTALL

• CONTACT LAURA HALE OR MIGUEL PARAZ:
  • SLACK CHANNEL: HTTPS://PARASPORTS.SLACK.COM
  • LAURA E-MAIL: LAURA@FANHISTORY.COM
  • MIGUEL E-MAIL: MPARAZ@GMAIL.COM
  • MIGUEL’S PARASPORT WIKIBASE TOOLS: HTTPS://GITHUB.COM/MPARAZ/PARASPORTS-DATA-IMPORT
MORE RESOURCES FOR YOUR OWN WIKIBASE INSTALL

- COMMUNITY RESOURCES:
  - GITHUB WIKIDATA TOOLKIT: HTTPS://GITHUB.COM/WIKIDATA/WIKIDATA-TOOLKIT
  - BLAZEGRAFH DOCUMENTATION: HTTPS://WIKI.BLAZEGRAFH.COM/WIKI/INDEX.PHP/MAIN_PAGE