Tales of doing Research with Video Game Fan Databases
A data-driven Approach

Nottingham, 21.08.2018
Peter Mühleder
Tracy Hoffmann
Cooperation between University Library Leipzig and Japanese Studies (Institute of East Asian Studies) of Leipzig University

Q: How can we use online data sources for research on (Japanese) video games?
Agenda

Video game fan databases as data sources
Linking data sources
Conceptual issues
Conclusion
Fan Databases

- Fans collect, organize and share an enormous amount of information about video games.
  e.g. Mobygames, GameFAQs, Wikia

- Highly specialized communities
- (Mostly) easily accessible
Fan Databases

- Metadata
  Company and developer credits, technical information, ...

- Discursive data
  User reviews, discussions, walkthroughs, ...

- Community practices
# Data Sources

Different sources provide different information

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Records</th>
<th>Language</th>
<th>Scope</th>
<th>Japanese Release Date(s)</th>
<th>Credits</th>
<th>Companies</th>
<th>Alternative Titles</th>
<th>Links to Knowledge Base</th>
<th>Walkthroughs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media Art DB</td>
<td>38.068</td>
<td>Jp</td>
<td>Japan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobygames</td>
<td>81.609</td>
<td>En</td>
<td>Worldwide</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GameFAQs</td>
<td>55.834</td>
<td>En</td>
<td>Worldwide</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

=> Data integration
Challenges

- No unique identifiers
- Datasets in different languages
- Heterogeneous data models

Q: How can we create links between these data sources?
Record linking

Linking based on game titles
=> Fan databases often feature alternative titles (in different languages)

=> Game title matching algorithm:
  - preselection based on platform
  - probabilistic ratio based on title similarity
  - extraction and comparison of numbers; subtitles
  - basis for machine learning model

https://github.com/diggr/diggrtoolbox

Result: Match probability $0 \leq x \leq 1.0$

[Mobygames entry X] $\leftarrow 0.9 \rightarrow$ [Media Art DB entry Y]
Linking results / clusters

Media Art DB -> Mobygames: 48 %
Media Art DB -> GameFAQs: 85 %
Linking results / clusters

Media Art DB -> Mobygames: 48 %
Media Art DB -> GameFAQs: 85 %
Linking results / clusters

Media Art DB -> Mobygames: 48 %
Media Art DB -> GameFAQs: 85 %

Visualization as network:

- Easy to identify wrong links
  Example
- Helps to identify problematic data
  Example
Linking results / clusters

Big cluster with ~ 9000 Datapoints (games)

Big franchises such as
Super Mario, Final Fantasy,
The Legend of Zelda, ...
Linking results / clusters

Big cluster

=> community detection algorithm

- 144 smaller clusters

- Mostly focused on a specific game/series
Linking clusters
Linking results / clusters

Problems:

- Missing links - “We don’t know what we don’t know”
- Link ratios are not very expressive
Behind the links

Conceptual Issues
“This Mobygames ID has a match with that Media Art DB ID”
“This Mobygames ID has a match with that Media Art DB ID”

“The release information of a Mobygames ID are the same as in this Media Art DB ID”

“This game in Mobygames is the same game as that Media Art DB ID”
Thoughts about links/matches/cluster

- Links have no semantics (at this state)
- Different goals and perspectives causes heterogeneous data models
- Researcher ask for a “game”
Thoughts about links/matches/cluster

- Links have no semantics (at this state)
- Different goals and perspectives causes heterogeneous data models
- Researcher ask for a “game”

Record ≠ Game

Mobygames  => New Edition  = New Record  
Media Art DB  => New Release  = New Record  
GameFAQs  => New Platform  = New Record
Conceptual entities of video games

- **Game**
  - **isDevelopedBy**
    - **Platform**
  - **isPublishedAs**
    - **Edition**
  - **isDistributedAs**
    - **Local Release**

<table>
<thead>
<tr>
<th>(Common) Game title</th>
<th>Nier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>XBox 360</td>
</tr>
<tr>
<td>Edition</td>
<td>NieR Gestalt (Origin game)</td>
</tr>
<tr>
<td>Release data</td>
<td>04/22/10</td>
</tr>
</tbody>
</table>
Tales of doing Research with Video Game Fan Databases

Graph:

- Game
  - hasPlatform
  - hasEdition
    - Mobgames
      - ID
    - hasRelease
      - Media Art DB
        - ID
Tales of doing Research with Video Game Fan Databases

Gamefaqs
Mobygames
Media Art DB
Video game data models

- clarify what you talking about
- no “one model to rule them all” -> depends on research question

Ongoing development
- add some semantics to the links (eg. belongsToSameSeries, sameTitleAs)
- create the game entity
Conclusion

By combining and using online video game databases we can

=> Build better research datasets
=> Build a video game reference dataset

But:
=> Automatic linking still has room for improvements
=> Conceptual model is required that can incorporate the different data models
=> It’s a long-term project
Thank you!

https://diggr.link/
https://github.com/diggr

licensed under Creative Commons Namensnennung 4.0 International Lizenz