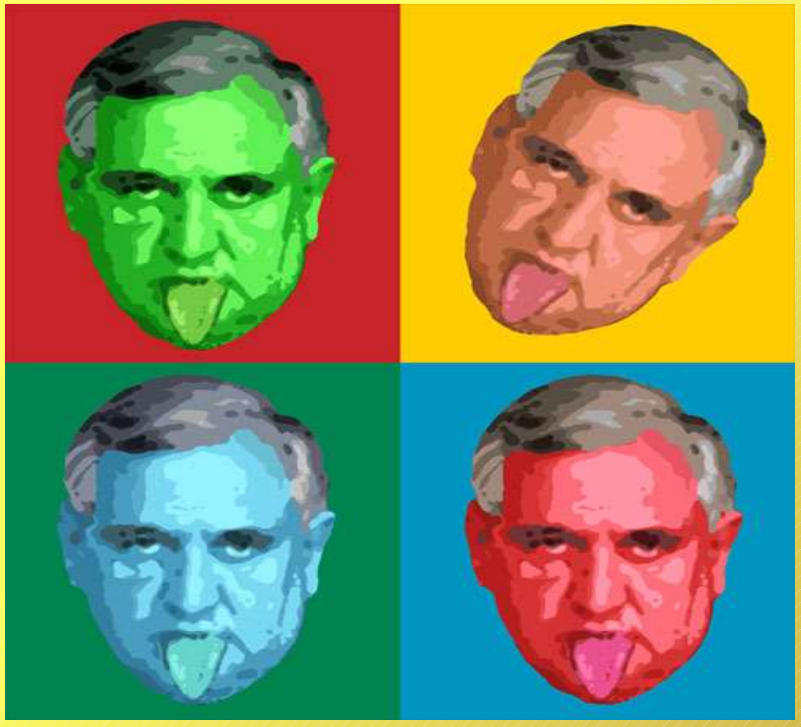
Developer Identification Methods for Integrated Data from Various Sources



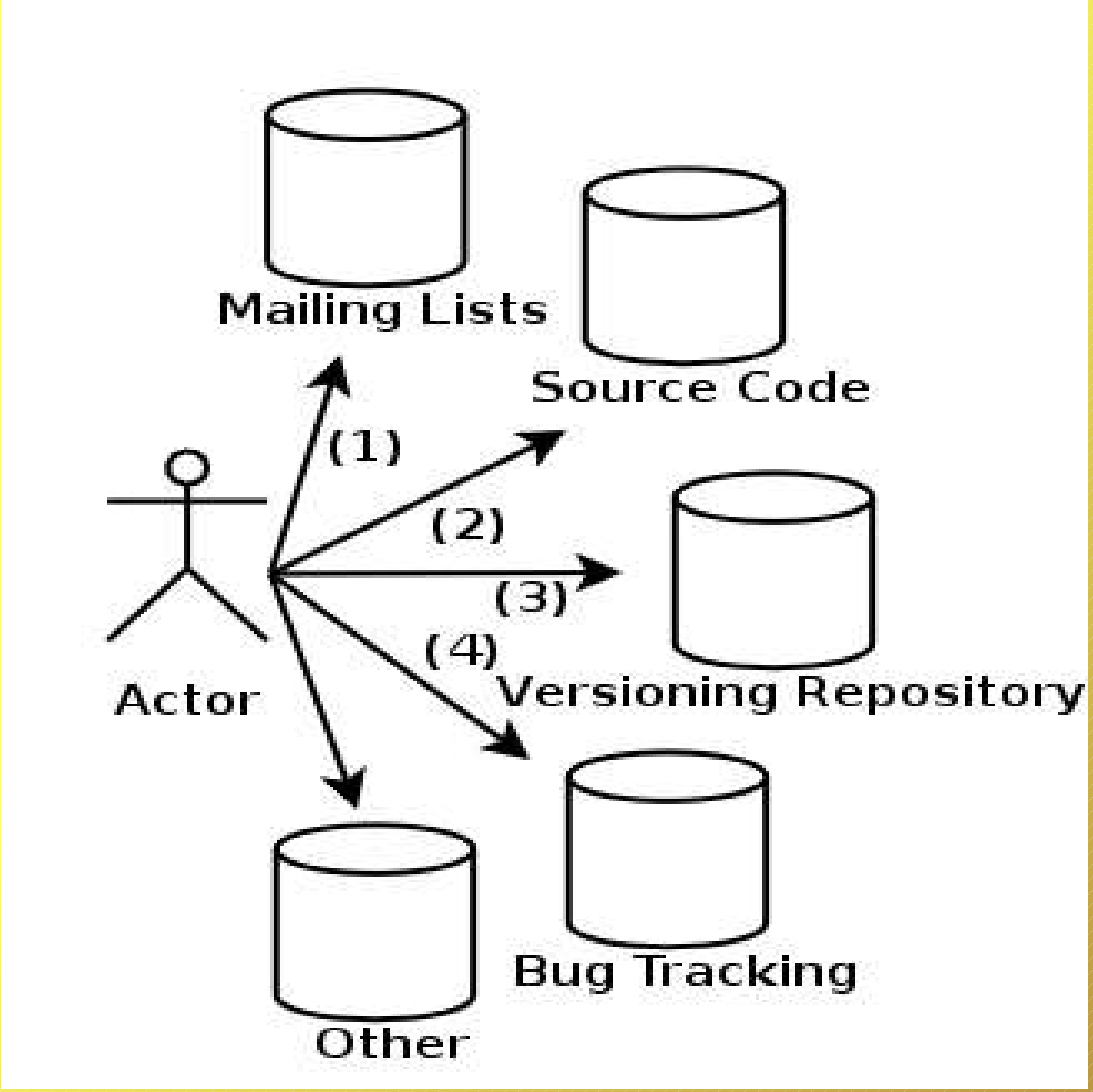
International Workshop on Mining Software Repositories (MSR 2005)
May 17th 2005, St. Luis, USA

Gregorio Robles, Jesús M. González-Barahona
Universidad Rey Juan Carlos – Madrid (Spain)
{grex,jgb}@gsyc.escet.urjc.es

Introduction

- Most MSR research has been performed on a single source of data
- When integrating data from various sources, we have to link artifacts:
  - Software artifacts: files, classes, functions...
  - Bug reports (bug report ids in CVS logs...)
  - Developers!

One developer, several identities



Sources and Identities (I)

- Developers have several ways of identifying in different tools (mostly this is tool-driven)
- There are primary identities and secondary ones
  - Primary are mandatory
    - For instance, in mailing lists you need an e-mail address
  - Secondary are redundant
    - For instance, in mailing list messages you can append your name to your e-mail address
- John Smith <john.smith@nowhere.com>

Sources and Identities (II)

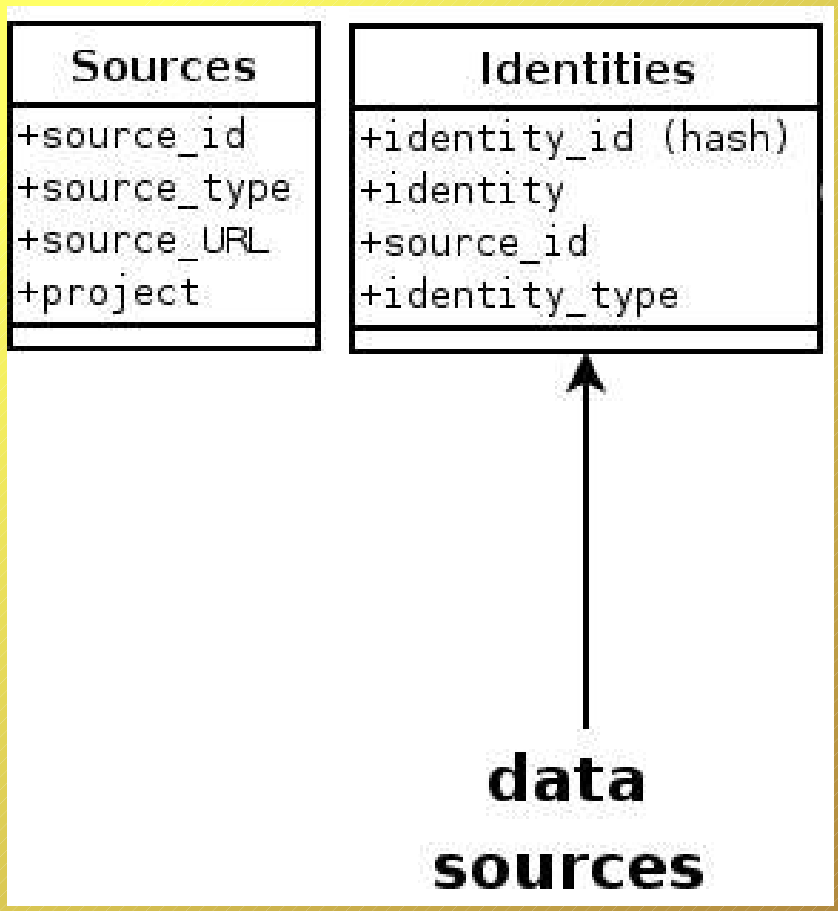
- Let's have a look at the primary identities:

	Source	Primary Identities	Tool
(1)	Mailing lists	username@example.com	MLStats
(2)	Source code	© Name Surname	pyTernity
(2)	Source code	© username@example.com	pyTernity
(2)	Source code	$id: username$	pyTernity
(3)	Versioning repository	username	CVSAnalY
(4)	Bug tracking system	username@example.com	BTSStats

- We can extract this information easily with current methodologies and tools

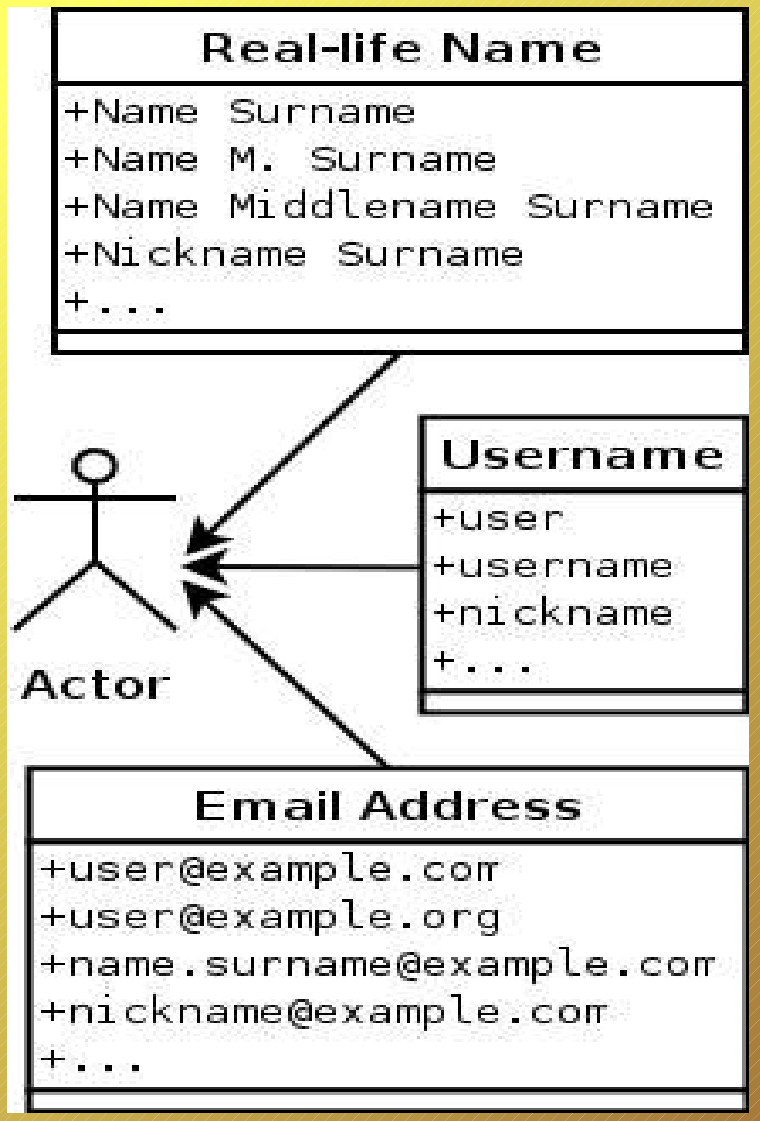
Finding Identities

- Directly from the sources (tools already exist for that)
- Hash as id
- Additional information about the data source



Identities for an actor

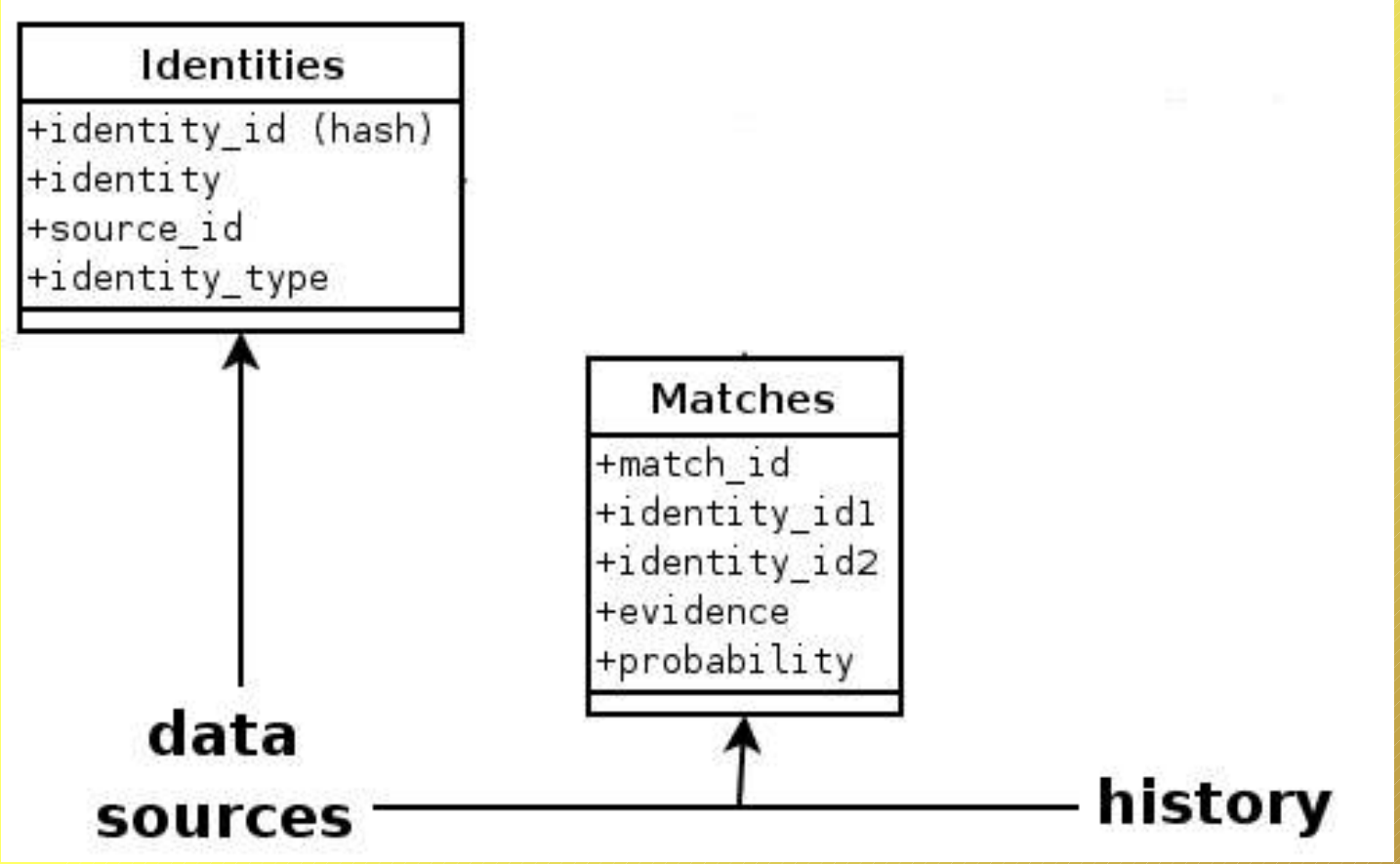
- We can classify identities in three groups:
- Notice that for each group we may have several entries even for the same developer!



Matching identities: Methods

- Directly from the data sources (redundancy)
  - Name Surname <username@example.com>
- Extracting 'real name' from e-mail (machine learning techniques exist for this)
  - name.surname@example.com
- Matching for username (e-mail, CVS)
  - Shrinking data to avoid many false positives
- Other sources: GPG Key rings, KDE CVS file, SF.net, Linux Credits, Google searches...

Finding Matches

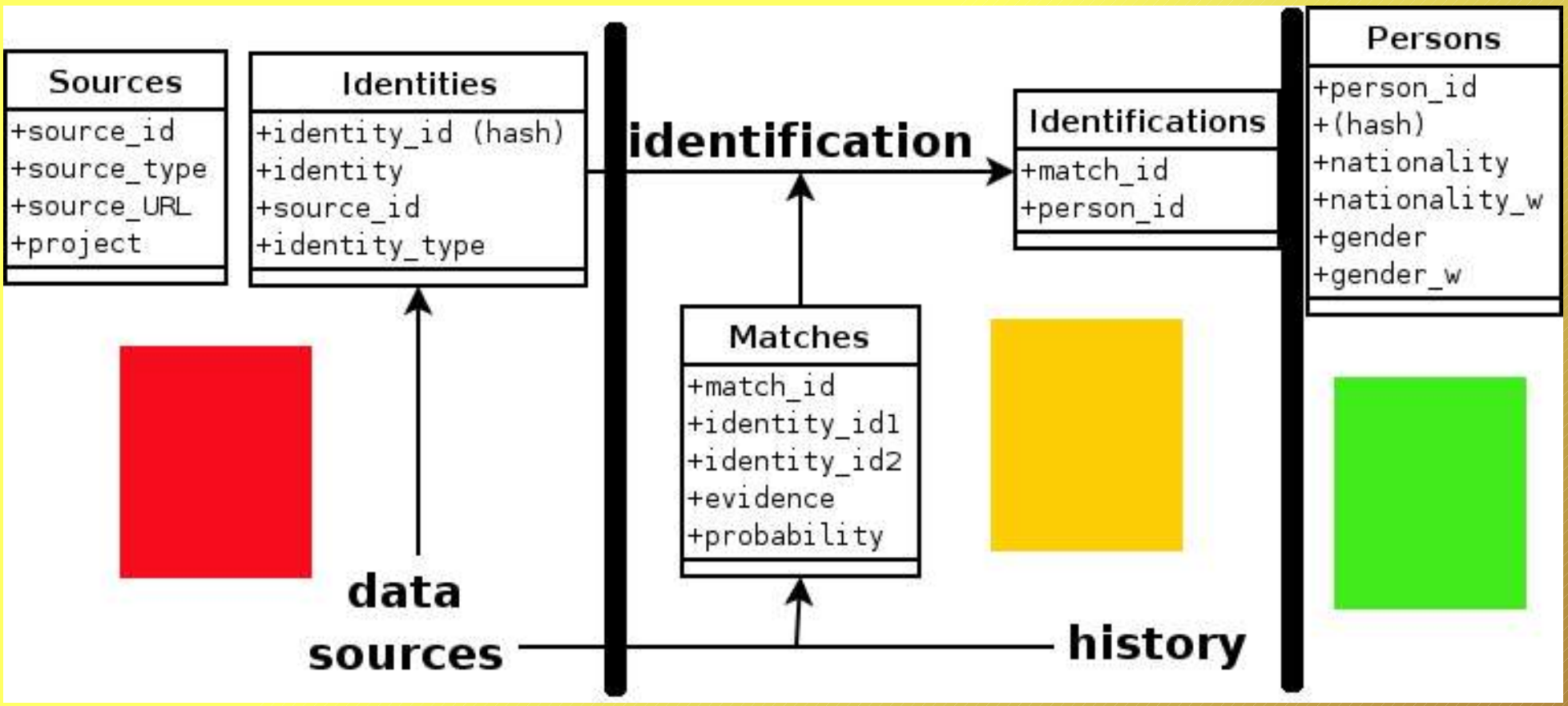


- Identities & Matches are fed from the data sources. Matches also allows historical info

Privacy Issues and Other Concerns

- We have (very) sensible data!
  - Not all the data set can be publicly available
- We would like to make studies reproducible
  - Data should be publicly available in an anonymized way to everybody
- We would like to share our data with other researchers; and matches should be enhanced
  - At least restricted access (available per request) for the matches table

Final Data Structure



(Notice distribution limitations through colors)

Case Study: GNOME

- Some stats:
  - Mailing lists: 464,953 messages from 36,399 distinct e-mail addresses
  - Bugs and comments on the Bug Tracking System: 123,739 bugs from 41,835 reporters and 382,271 comments from 10,257 posters
  - CVS writers: around 2,000,000 from 1,067 commiters
  - (Source code has not been scanned)

Some (preliminary) results

- After the identification process:
  - 108,170 distinct identities for all sources
  - 40,003 distinct author matches
  - Identified 34,648 different persons
- Matches table has been manually inspected in part and matches are consistent
- We have now a third of the previous (virtual) population

Conclusions

- Developers have different identities for tools used in libre software development
- We can track developers through different sources using redundant data that is available
- We have made a proposal for preserving anonymity while at the same time sharing data among researchers

Additional Information

- From the TLD of the e-mail addresses we may infer nationality (.de -> German, .fr -> France)
  - What should we do with generic TLDs?
- From the whois information of the domain we may infer nationality (barrapunto.com -> Spain)
- From nationality and name we may infer gender (Andrea is a male name in Italy, not in Germany)
- All this information is based on heuristics and is fuzzy (additional field for it)

References

- [Koch00]
- [Mockus02]
- [Robles03]
- [FLOSS02]