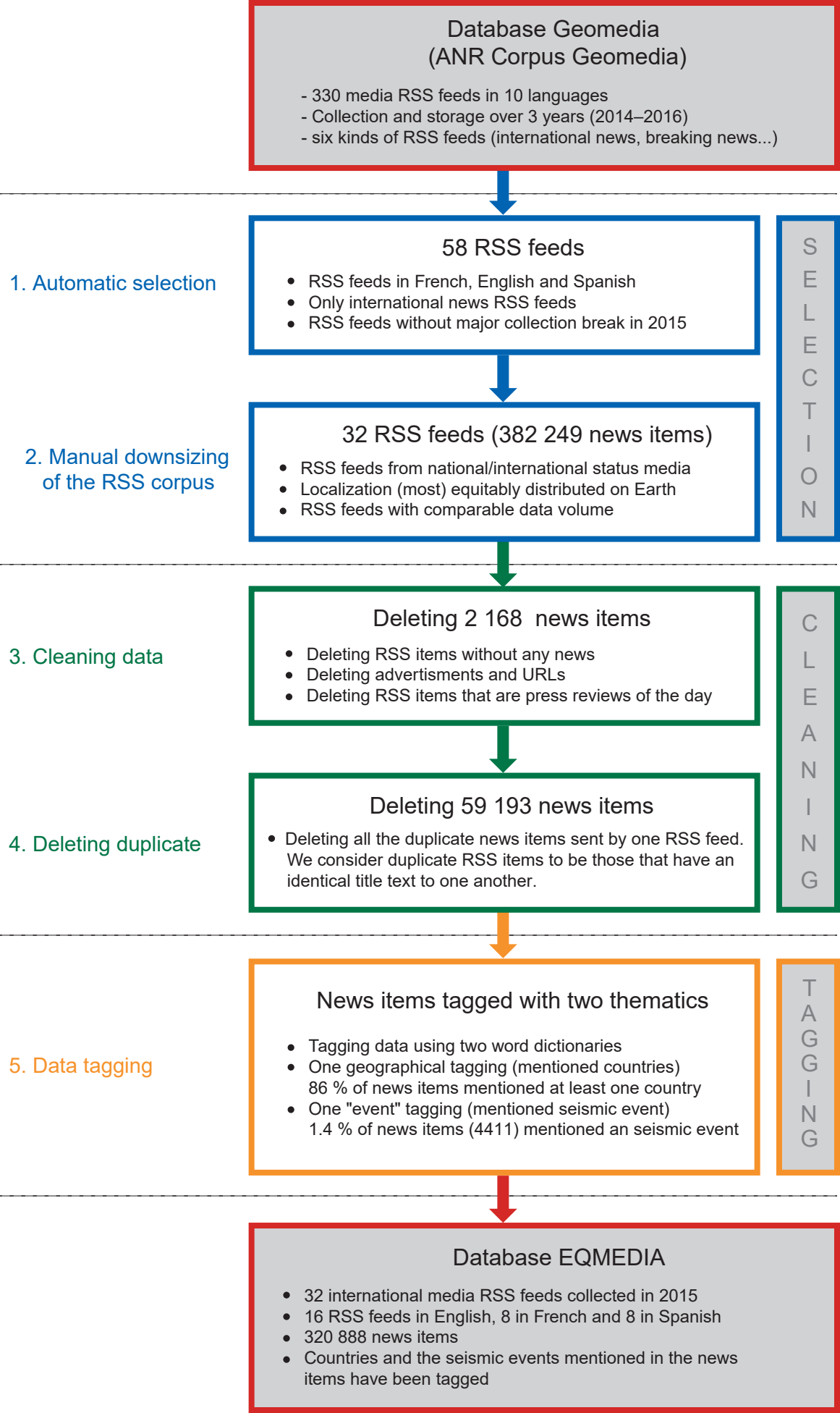
Database Geomedia
(ANR Corpus Geomedia)
- 330 media RSS feeds in 10 languages
- Collection and storage over 3 years (2014–2016)
- six kinds of RSS feeds (international news, breaking news...)
1. Automatic selection
58 RSS feeds
RSS feeds in French, English and Spanish
Only international news RSS feeds
RSS feeds without major collection break in 2015
2. Manual downsizing of the RSS corpus
32 RSS feeds (382 249 news items)
RSS feeds from national/international status media
Localization (most) equitably distributed on Earth
RSS feeds with comparable data volume
SELECTION
3. Cleaning data
Deleting 2 168 news items
Deleting RSS items without any news
Deleting advertisments and URLs
Deleting RSS items that are press reviews of the day
4. Deleting duplicate
Deleting 59 193 news items
Deleting all the duplicate news items sent by one RSS feed. We consider duplicate RSS items to be those that have an identical title text to one another.
CLEANING
5. Data tagging
News items tagged with two thematics
Tagging data using two word dictionaries
One geographical tagging (mentioned countries)
86 % of news items mentioned at least one country
One "event" tagging (mentioned seismic event)
1.4 % of news items (4411) mentioned an seismic event
TAGGING
Database EQMEDIA
32 international media RSS feeds collected in 2015
16 RSS feeds in English, 8 in French and 8 in Spanish
320 888 news items
Countries and the seismic events mentioned in the news items have been tagged