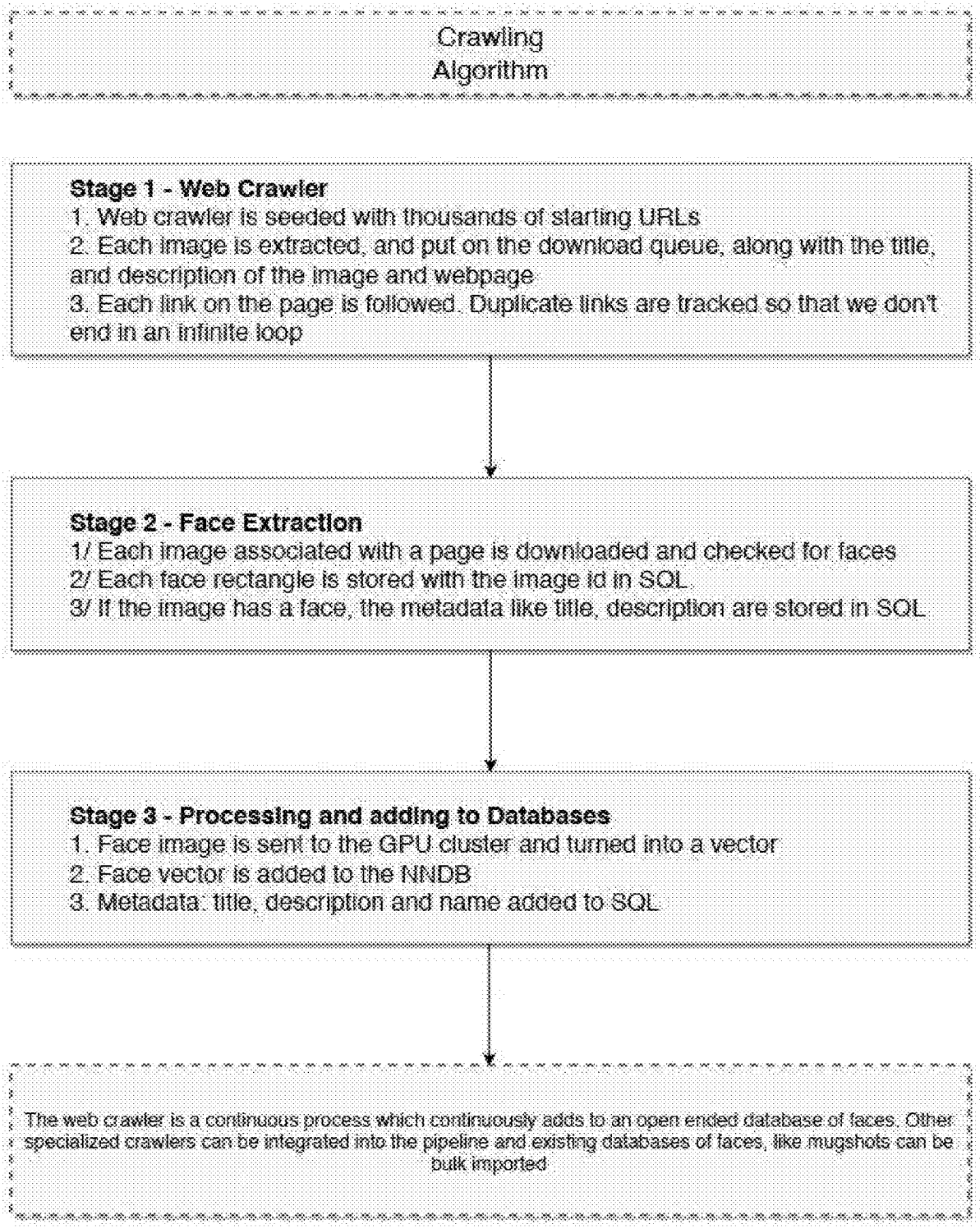
Crawling
Algorithm
Stage 1 - Web Crawler
1. Web crawler is seeded with thousands of starting URLs
2. Each image is extracted, and put on the download queue, along with the title, and description of the image and webpage
3. Each link on the page is followed. Duplicate links are tracked so that we don't end in an infinite loop
Stage 2 - Face Extraction
1/ Each image associated with a page is downloaded and checked for faces
2/ Each face rectangle is stored with the image id in SQL.
3/ If the image has a face, the metadata like title, description are stored in SQL
Stage 3 - Processing and adding to Databases
1. Face image is sent to the GPU cluster and turned into a vector
2. Face vector is added to the NNDB
3. Metadata: title, description and name added to SQL
The web crawler is a continuous process which continuously adds to an open ended database of faces. Other specialized crawlers can be integrated into the pipeline and existing databases of faces, like mugshots can be bulk imported

FIG. 3