Ensuring content flexibility

Challenges in newspaper media monitoring

Tim Saulter, Product Marketing Manager
Newspaper Licensing Agency

About the NLA database

- **150 UK newspapers**

- **XML and PDF (page and 'clips')**

- **16,000 new articles per day**

- **Core: Wholesale to UK media monitors**

- **Spin-off services**
  - Text aggregator supply e.g. Factiva, Lexis Nexis
  - Mobile devices e.g. Kindle and other eReaders
  - Library services e.g. journalists (ClipShare), schools (Newspapers for Schools)

Key drivers

- **Large daily updates, lots of (relatively) small articles**

- **Timeliness of update is imperative**

- **Exact concordance with published content (all editions)**

- **Easy to distribute outbound to multiple recipients/applications**

- **Secure and usage-monitored environment**

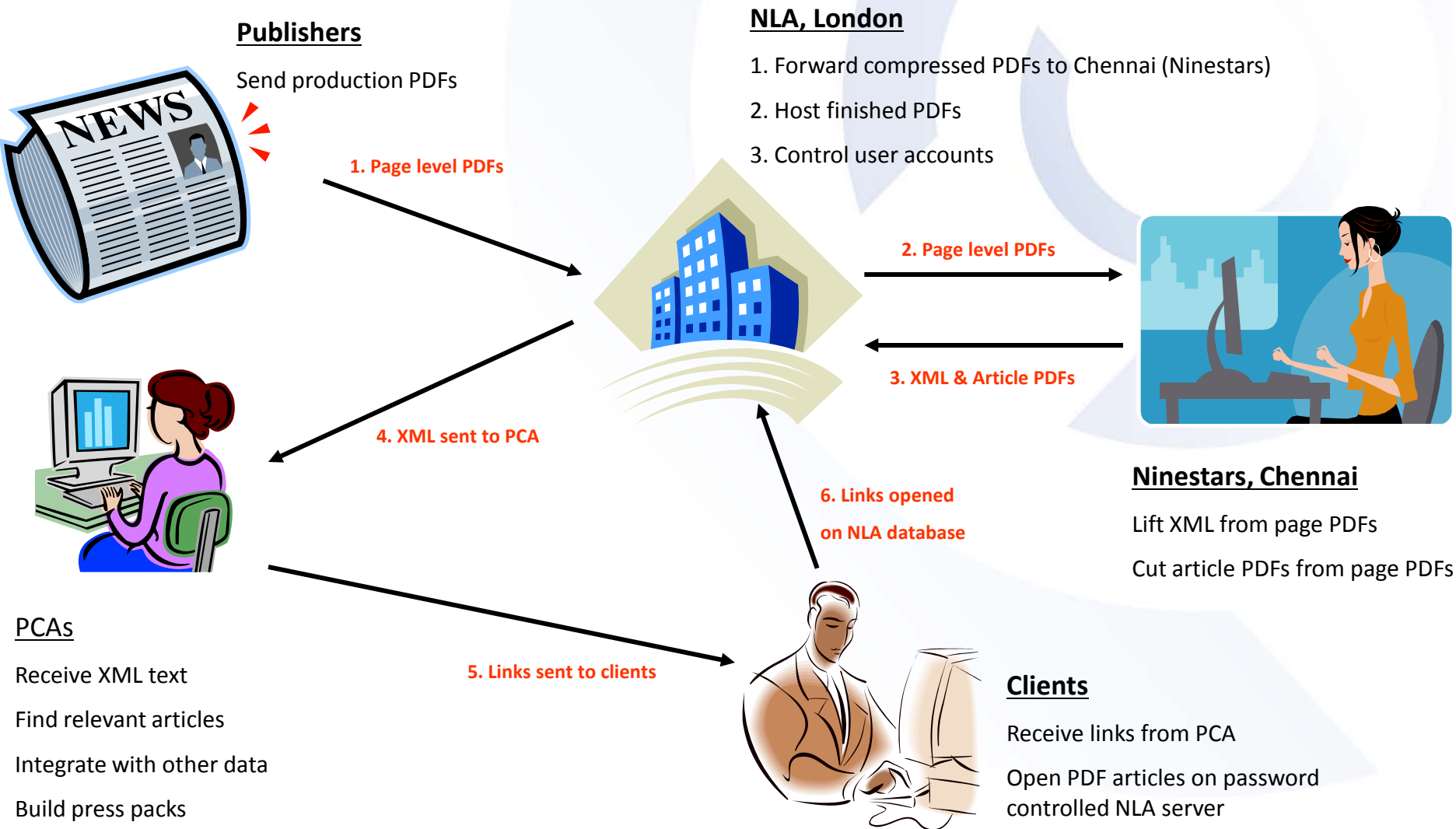**Note: We do no categorisation or semantic tagging**

Consistent XML is utopia?

- **XML = flexibility, but...**

- **"Your XML" does not always = "My XML"**

- **Standards crumble under business pressure**

- **Core of newspaper business is paper, not XML**

- **So we cannot use the source data**

How do we do it?

Publishers

Send production PDFs

NLA, London

1. Forward compressed PDFs to Chennai (Ninestars)

2. Host finished PDFs

3. Control user accounts

1. Page level PDFs

2. Page level PDFs

3. XML & Article PDFs

4. XML sent to PCA

6. Links opened

on NLA database

Ninestars, Chennai

Lift XML from page PDFs

Cut article PDFs from page PDFs

PCAs

Receive XML text

Find relevant articles

Integrate with other data

Build press packs

5. Links sent to clients

Clients

Receive links from PCA

Open PDF articles on password controlled NLA server

Why do it 'the hard way'?

- **Publisher does not need to do anything new**

- **We get XML to our own specification**

- **Use a good outsource supplier = quick & reliable**

- **Very reliable; our KPIs are for 99% timeliness (by 0400) and accuracy**

- **Downside.... Expensive to scale**

The future of print processing

- **Taking it from the source (CMS) allows scale, but...**

- **Standards still don't exist**
    - Publish our own standards
    - Work with publishers and CMS providers
    - See who follows

- **Physical storage remains a high proportion of costs**
    - Cloud storage solutions?

- **May remain a pipedream; push for semi-automation instead**

NLA and the web

- **Represents a new challenge: realtime, versioning, transience**

- **We are going direct from CMS – aim to beat scraping**

- **Creating XML feed and secure permanent archive**

- **It works but is hard, e.g. CDATA**

- **Lack of standards is now the blocker to growth**

- **High staffing costs**

So what?

- **Good quality aggregated XML opens many doors**

- **But business realities can impact the quality from source**

- **We had to acknowledge that repurposing content is a distant second**

- **Re-aligning a traditional publishing business is tough and slow**

- **But anything can be worked around – and companies to help**