Research Source Download System

Overview

This system systematically downloads and organizes research materials from your BBNJ and CHM reference documents for offline access during writing.

Setup & Usage

1. Extract URLs from markdown files

```
node extract-urls.js
```

bas

This creates:

- `sources/urls-extracted.json` - All URLs in JSON format
- `sources/urls-report.md` - Human-readable report
- `sources/download.sh` - Basic download script

2. Download PDFs and prepare web content

```
node download-sources.js
```

bas

Downloads PDFs directly and creates metadata files for web content that needs scraping.

3. Scrape web articles

```
node web-scraper.js
```

bas

Fetches HTML content and converts it to markdown for easy reading.

4. Generate index

```
node generate-index.js
```

bas

Creates comprehensive index of all downloaded materials.

Current Status

✅ **Completed:**

- Extracted 222 unique URLs from your reference documents
- Downloaded 6 PDF documents (8MB total)
- Scraped 9 web articles to markdown format

- Created organized folder structure
- Generated comprehensive index

⏳ **Pending Enhancements:**

1. **DOI Resolution**: Use Unpaywall API to find open-access versions of academic papers
2. **PDF to Markdown**: Convert PDFs to searchable markdown text
3. **Advanced Web Scraping**: Use Playwright for JavaScript-heavy sites
4. **Batch Processing**: Process remaining ~200 URLs

Downloaded Content Structure

```
sources/
├── pdf/          # PDF documents
├── article/      # Web articles (markdown)
├── academic/     # Academic papers
├── official/     # UN/government documents
├── metadata/     # Processing metadata
├── INDEX.md      # Human-readable index
└── index.json    # Machine-readable index
```

Key Files Successfully Downloaded

PDFs

- IUCN BBNJ Treaty Policy Brief
- ISA Contribution to BBNJ Objectives
- High Seas Alliance PrepCom Brief

Web Content

- ECO Law Blog on hydrothermal vents
- IDDRI policy briefs
- Inside EU Life Sciences analysis
- NCLOS Blog on CITES-BBNJ interface
- Opinio Juris on MGR governance

Next Steps for Full Coverage

To download the remaining ~200 sources:

1. **Install additional dependencies** (optional):

```bash
npm install playwright puppeteer-core pdf-parse turndown
```

2. **Run enhanced scraper** with retry logic for failed URLs

3. **Use DOI resolver** for academic papers:

    - Implement Unpaywall API integration

    - Check institutional repositories
    - Try Sci-Hub alternatives (where legal)
4. **Convert PDFs to markdown** for better integration:

```bash
npm install pdf-to-markdown
```

Using Downloaded Sources in Your Writing

The downloaded materials are now available locally in the `sources/` directory. When writing your research:

1. Reference the `sources/INDEX.md` for a complete list
2. Use local file paths instead of URLs
3. All content is in markdown format for easy integration
4. PDFs are preserved in original format for citations

Manual Download for Critical Sources

For sources that failed automatic download, you can manually download and add them to the appropriate folder. The most important ones to prioritize:

1. UN official documents (UNEP, UNESCO)
2. Nature editorial (may require institutional access)
3. Frontiers articles (should be open access)
4. EJIL:Talk! blog posts (should be accessible)

Quick Commands Reference

```bash
# Run complete pipeline
node extract-urls.js && node download-sources.js && node web-scraper.js && node generate-index.js

# View results
cat sources/INDEX.md

# Count successfully downloaded files
ls -la sources/*/ | wc -l
```