数据确权方案讨论

明确的职责

- 数据权属注册：在链上记录权属信息，确定数据源头，构建数据和权属之间的联系。
- 权属流转控制：数据的传播控制，访问控制

假设的前提

- 存在至少半可信的第三方：用于对数据的所有权进行认证背书
- 第三方对数据的原创性检验仅限于检测是否存在数据已被注册或部分数据重复注册的请客，而不设计数据的抄袭等客观难判断的情况（是否合法、原创由第三方保证，直接默认上链信息是可信的）

可用的工具

链上权属和链下数据的链接：

- 哈希函数：（链上：摘要；链下：原始数据）
  - 优点：快速、高效；
  - 缺点：鲁棒性差；
- 分块哈希、模糊哈希：（链上：模糊摘要；链下：原始数据）
  - 优点：快速、高效；
  - 缺点：鲁棒性稍差，对恶意用户失效；
- 布隆过滤器：（不应该在链上，只能在第三方辅助使用）
  - 优点：快速匹配，高效。
  - 缺点：只能确定数据是否一定不存在，只能算初步验证；鲁棒性稍差，对恶意用户失效；
- 水印（非零水印）：（链上：具体的水印内容；链下：嵌入水印后的数据）
  - 优点：鲁棒性强，能够抵御一定恶意用户；功能相对更多；
  - 缺点：侵入式，需要在原始文件中嵌入信息；相对耗时，效率低；
- 深度学习（特征提取）：（链上：数据特征；链下：原始数据）
  - 优点：鲁棒性强，能够抵御一定恶意用户；
  - 缺点：耗时高、效率低；需要编码器、解码器；

目前的想法

通过比对针对“链接链上权属与链下数据”的各类方法，可以看出：

- 一部分方法适用于快速建立链接（弱链接，匹配到一定是；不匹配到不一定不是（恶意用户情况））
- 一部分方法适用于鲁棒地仲裁（强链接，只要成功提取到完整信息，就能够映射到具体权属）

初步方案想法：

- 两者结合，链上记录（数据拥有者等元数据，数据模糊摘要），链下在数据中嵌入有关数据权属的水印信息。
    - 弱链接用于快速提取数据的模糊摘要，找到链上的权属信息。
    - 强链接用于仲裁数据的非法传播，防止恶意用户通过更改文件特征影响链接。
- 设定数据注册中心，解决数据的权属证明的生成以及水印的嵌入。

考虑点：

- 为什么不直接使用水印：
    - 为抵御恶意用户，用户不应该能过直接从嵌入水印文件中提取原始文件或水印（盲水印），应采用只有获得部分水印信息或原始文件才能获得原始文件的水印方式（非盲水印、半盲水印）。所以水印信息不能由用户自行提取，也不宜直接存储链上。
    - 采用非盲水印或半盲水印，只能与第三方交互之后或由第三方地区才能提取水印，以匹配到链上存储，这样提取和嵌入需要时间及资源消耗，存在单点故障问题。
- 为什么不直接使用模糊哈希等特征提取技术：
    - 主要原因是其无法有效地抵御恶意用户对文件内容的部分修改，以至于找到非法传播的文件无法正确匹配到链上的权属信息。

考虑的衍生问题

- 数据权属存储链上的隐私问题：数据权属信息本身是否敏感，数据权属的流转是否敏感。
- 数据权属的链上索引问题：拿到数据模糊摘要后如何在链上快速确定流转路径，而非遍历区块链。
- 数据流转控制问题：也就是数据确定权属之后的权属变更问题，涉及数据的传播控制、访问控制。（之前认为的确权第二部分）
- 数据确权中身份系统问题：即数据权属注册、转移过程中各方身份问题，涉及密钥协商、访问控制、隐私问题。